SHARP LIMITS AND EFFICIENT RECOVERY OF LATENT MEMBERSHIPS IN RANDOM-DESIGN REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixed regression models capture population heterogeneity through distinct regression relationships across latent groups. When group memberships are scientifically or practically meaningful, recovering them becomes a fundamental inferential problem. Yet the statistical limits of membership recovery remain much less understood than those of estimating the regression components. For the two-component Gaussian random-design model with $p \lesssim n$, we show that the minimax mis-clustering risk is of order $\min(1, r)$ for noise-to-signal ratio r, while exact recovery is possible if and only if $nr \to 0$. These limits arise from an intrinsic ambiguity induced by the random regression margin, whose mass near zero determines the mis-clustering scale, while the minimum realized margin across the sample governs exact recovery. We further develop an efficient procedure attaining both limits, in which a bounded-residual spectral direction is combined with the pooled residual to form an initial partition, followed by hard least-squares iterations that geometrically contract a margin-weighted Hamming loss.

1 INTRODUCTION

Mixed regression models have long been used to represent population heterogeneity through regression relationships that differ across latent groups (Quandt & Ramsey, 1978). In many applications, the latent memberships themselves are scientifically or practically important, supporting segmentation in marketing (DeSarbo & Cron, 1988), informing treatment decisions in precision medicine (Ma et al., 2020), and providing the grouping structure for subsequent estimation and inference in economic panel analysis (Su et al., 2016). Yet much of the statistical theory for mixed regression has focused on estimation of the regression components, addressing both minimax rates and computationally efficient recovery (Chen et al., 2014; Kwon et al., 2021; Tan & Venkataramanan, 2023). By contrast, the statistical limits of recovering the realized memberships themselves remain much less understood.

We study this recovery problem in the two-component random-design regression model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^*_{z_i^*} + \epsilon_i, \qquad \mathbf{x}_i \overset{\text{iid}}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma}), \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad i = 1, \dots, n,$$

where the memberships $z_i^* \in \{+1, -1\}$ are fixed but unknown, with each group containing at least $c_0 n$ observations for some fixed $c_0 \in (0, 1/2)$. Here the noise is independent of the design.

We write $\boldsymbol{\beta}^*_+, \boldsymbol{\beta}^*_- \in \mathbb{R}^p$ for the two regression vectors, using $+$ and $-$ as shorthand subscripts for $+1$ and -1, respectively. The covariance $\boldsymbol{\Sigma} \succ 0$ and noise level $\sigma \geq 0$ are unknown. For $\mathbf{h} \in \mathbb{R}^p$, let $\|\mathbf{h}\|_{\boldsymbol{\Sigma}} := (\mathbf{h}^\top \boldsymbol{\Sigma} \mathbf{h})^{1/2}$. Set

$$\boldsymbol{\theta} := \boldsymbol{\beta}^*_+ - \boldsymbol{\beta}^*_-, \qquad \Lambda := \left(\mathbb{E}[(\mathbf{x}_i^\top (\boldsymbol{\beta}^*_+ - \boldsymbol{\beta}^*_-))^2]\right)^{1/2} = \|\boldsymbol{\theta}\|_{\boldsymbol{\Sigma}}, \qquad r := \frac{\sigma}{\Lambda},$$

and we assume $\Lambda > 0$ throughout. The regression components are shared across observations, while each observation has its own membership to be recovered.

The central difficulty in recovering z_i^* is that, conditional on $\mathbf{x}_i$, the two possible response means are separated by the observation-specific margin $|\mathbf{x}_i^\top \boldsymbol{\theta}|$, which can be arbitrarily small. Under Gaussian design,

$$\frac{|\mathbf{x}_i^\top \boldsymbol{\theta}|}{\Lambda} \overset{d}{=} |G|, \qquad G \sim N(0, 1),$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

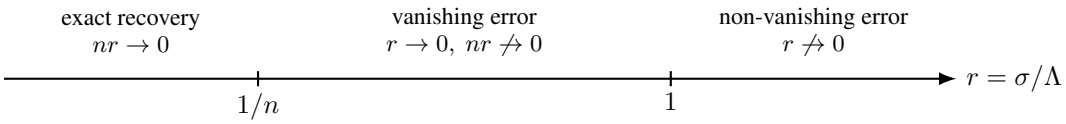


Figure 1: Schematic recovery regimes as a function of $r = \sigma/\Lambda$.

and the density of $|G|$ is positive at zero. Hence, when r is small, an order-r fraction of observations have separation at or below the noise scale. We show that this small-margin probability determines the minimax mis-clustering scale $\min(1, r)$. For exact recovery, the smallest normalized margin is of order $1/n$. We show that exact recovery is possible if and only if $nr \to 0$. Figure 1 summarizes these regimes.

These recovery scales are intrinsic to the memberships rather than a consequence of estimating the regression components. Even when the regression vectors are known, the symmetric oracle problem has single-membership Bayes error $q(r) = \pi^{-1} \arctan(2r)$; see Section 2. This source of ambiguity differs from that in Gaussian location mixtures, where component separation is fixed across observations and ambiguity is governed by Gaussian tails (Ndaoud, 2022; Chen & Yang, 2021). Profile least squares attains the same mis-clustering scale, up to an exponentially small remainder, by searching over candidate membership vectors and fitting the two regressions within each candidate partition.

This raises the natural computational question of whether statistically optimal membership recovery can be achieved efficiently. Profile least squares requires a combinatorial search over exponentially many candidate membership vectors. We develop a polynomial-time procedure that attains both recovery scales. The key theoretical ingredient is a uniform local analysis of hard least squares showing that, once the fitted regression components enter a fixed neighborhood of the truth, the margin-weighted Hamming loss satisfies a geometric contraction bound with a remainder small enough to attain both recovery scales. To initialize, we estimate the separation direction using a bounded residual spectral statistic. The estimated direction, together with the pooled residuals, defines a preliminary partition for which the hard least-squares refit lies in this neighborhood.

1.1 Main results

We measure mis-clustering error up to global label switching by

$$\bar{h}(\widehat{\mathbf{z}}, \mathbf{z}) := \frac{1}{n} \min_{\eta \in \{+1,-1\}} \sum_{i=1}^{n} \mathbf{I}(\widehat{z}_i \neq \eta z_i).$$

For fixed c_0 and (Λ, σ), let

$$\Theta_{c_0}(n, p; \Lambda, \sigma) := \left\{ (\mathbf{z}, \boldsymbol{\beta}_+, \boldsymbol{\beta}_-, \boldsymbol{\Sigma}) : \begin{array}{l} \mathbf{z} \in \{+1,-1\}^n, \quad \min_{b \in \{+1,-1\}} |\{i : z_i = b\}| \geq c_0 n \\ \boldsymbol{\Sigma} \succ 0, \quad \|\boldsymbol{\beta}_+ - \boldsymbol{\beta}_-\|_{\boldsymbol{\Sigma}} = \Lambda \end{array} \right\}.$$

The minimax mis-clustering and exact-recovery risks are

$$R_H^*(n, p, \Lambda, \sigma) := \inf_{\widehat{\mathbf{z}}} \sup_{\vartheta \in \Theta_{c_0}(n,p;\Lambda,\sigma)} \mathbb{E}_\vartheta \bar{h}(\widehat{\mathbf{z}}, \mathbf{z}),$$
$$R_{\mathrm{ex}}^*(n, p, \Lambda, \sigma) := \inf_{\widehat{\mathbf{z}}} \sup_{\vartheta \in \Theta_{c_0}(n,p;\Lambda,\sigma)} \mathbb{P}_\vartheta(\widehat{\mathbf{z}} \notin \{\mathbf{z}, -\mathbf{z}\}).$$

Theorem 1.1 (Information-theoretic limits)**.** *There are constants $C_1, c_1, \kappa_1 > 0$, depending only on c_0, such that, for all sufficiently large n with $p \leq \kappa_1 n$,*

$$c_1 \min(1, r) \leq R_H^*(n, p, \Lambda, \sigma) \leq C_1 \min(1, r) + C_1 e^{-c_1 n}.$$

Along any sequence satisfying $p \leq \kappa_1 n$,

$$R_{\mathrm{ex}}^*(n, p, \Lambda, \sigma) \to 0 \quad \Longleftrightarrow \quad nr \to 0.$$

The recovery algorithm in Section 3 attains these limits. For $r > 0$ and $t \geq 1$, define the finite-sample scale

$$\Psi_{n,r}(t) := r + \sqrt{\frac{rt}{n}} + \frac{t/n}{\log(e + t/(nr))}.$$

Theorem 1.2 (Efficient recovery)**.** *Suppose $r > 0$. There are constants $C_2, c_2, \kappa_2 > 0$, depending only on c_0, such that, whenever $p + \log(en) \le \kappa_2 n$ and $K \ge \lceil 5\log(1/r) \rceil$, the estimator in Algorithm 3.1 satisfies, uniformly over $\Theta_{c_0}(n, p; \Lambda, \sigma)$,*

$$\mathbb{E}\bar{h}(\widehat{\mathbf{z}}, \mathbf{z}^*) \le C_2 \min(1, r) + C_2 e^{-c_2 n},$$
$$\mathbb{P}(\widehat{\mathbf{z}} \notin \{\mathbf{z}^*, -\mathbf{z}^*\}) \le C_2 n r + C_2 e^{-c_2 n}.$$

Moreover, for every $t \ge 1$ satisfying $p + t + \log(en) \le \kappa_2 n$,

$$\mathbb{P}\Big(\bar{h}(\widehat{\mathbf{z}}, \mathbf{z}^*) > C_2 \Psi_{n,r}(t)\Big) \le C_2 e^{-t}.$$

The same proof strategy also yields an exact-recovery guarantee in the noiseless setting.

Corollary 1.3 (Noiseless recovery)**.** *There are constants $C, c, \kappa > 0$, depending only on c_0, such that, when $\sigma = 0$, $p + \log(en) \le \kappa n$ and $K \ge C \log n$, the estimator in Algorithm 3.1 satisfies, uniformly over $\Theta_{c_0}(n, p; \Lambda, 0)$,*

$$\mathbb{P}(\widehat{\mathbf{z}} \notin \{\mathbf{z}^*, -\mathbf{z}^*\}) \le C n^{-c}.$$

Moreover, if $K \ge Cn$, the failure probability can be improved to Ce^{-cn}.

1.2 Related literature

Mixed regression. Regression models with latent group structure have long been studied through switching, clusterwise, and finite-mixture formulations (Quandt & Ramsey, 1978; Späth, 1979; DeSarbo & Cron, 1988; De Veaux, 1989; Hennig, 2000). In clusterwise, subgroup, and latent group panel regression, the latent partition is often itself an object of statistical interest, alongside the group-specific regression parameters (Park et al., 2017; Ma & Huang, 2017; Ma et al., 2020; Bonhomme & Manresa, 2015; Su et al., 2016). By contrast, modern theory for mixed linear regression has largely focused on the statistical limits and computationally efficient recovery of the regression components (Chaganty & Liang, 2013; Chen et al., 2014; Yi et al., 2014; Kwon et al., 2021; Tan & Venkataramanan, 2023). In particular, exact parameter recovery and fast local convergence have been established in noiseless mixed linear regression (Yi et al., 2014; Zhong et al., 2016; Ghosh & Ramchandran, 2020).

To our knowledge, theory for recovering the realized memberships is comparatively limited, with exact-recovery guarantees available under noiseless well-separated conditions (Hand & Joshi, 2018) or, for sparse models, under a uniform positive squared-residual gap for every observation (Barik & Honorio, 2022). Against this background, we take the realized membership vector as the primary object of interest and study its sharp recovery limits under noisy Gaussian random design, together with their efficient attainment.

Clustering and label recovery. Our problem is also related to a broad literature on cluster-label recovery. For finite-mixture clustering, sharp mis-clustering rates, exact-recovery thresholds, and computationally efficient procedures have been established under Gaussian and sub-Gaussian models (Azizyan et al., 2013; Lu & Zhou, 2016; Fei & Chen, 2018; Löffler et al., 2021; Chen & Yang, 2021; Ndaoud, 2022). More generally, Gao & Zhang (2022) develop a unified theory for iterative recovery of discrete latent structures by alternating between parameter estimation and label updates. A key distinction from the settings above is that, in our regression problem, the separation relevant for membership recovery is an observation-specific random margin with nonnegligible probability mass near zero.

Margin phenomena. The dependence of recovery difficulty on the distribution of the random margin is closely related to margin conditions in classification, where the probability mass near the decision boundary determines achievable statistical rates (Mammen & Tsybakov, 1999; Tsybakov, 2004; Massart & Nédélec, 2006; Audibert & Tsybakov, 2007). In our setting, the fundamental difficulty likewise comes from observations whose random regression margins lie near zero. Unlike supervised classification, however, the memberships are latent and the regression components determining these margins are themselves unknown. Interestingly, our results show that estimating the regression components incurs no additional statistical cost at the sharp membership-recovery scale, since the same scale already arises in the oracle problem with known components.

The remainder of the paper establishes the information-theoretic limits, presents the recovery algorithm, analyzes its local contraction and initialization, and reports numerical experiments before concluding with a discussion.

2 INFORMATION-THEORETIC LIMITS

For $\mathbf{z} \in \{+1,-1\}^n$, let $I_+(\mathbf{z}) := \{i : z_i = +1\}, I_-(\mathbf{z}) := \{i : z_i = -1\}$, and denote

$$\mathrm{RSS}(\mathbf{z}) = \min_{\boldsymbol{\beta}_+,\boldsymbol{\beta}_- \in \mathbb{R}^p} \Big(\sum_{i \in I_+(\mathbf{z})} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_+)^2 + \sum_{i \in I_-(\mathbf{z})} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_-)^2 \Big).$$

The profile estimator is defined as

$$\widehat{\mathbf{z}}^{\mathrm{prof}} \in \operatorname*{arg\,min}_{\mathbf{z} \in \mathcal{T}_{c_0/2}} \mathrm{RSS}(\mathbf{z}),$$

where $\mathcal{T}_{c_0/2} = \big\{\mathbf{z} \in \{+1,-1\}^n : |I_+(\mathbf{z})|, |I_-(\mathbf{z})| \geq c_0 n/2\big\}$. This set contains the true membership vector while imposing a mild balance constraint on candidate partitions.

Proposition 2.1 (Profile upper bound)**.** *There exist constants $C, c, \kappa > 0$, depending only on c_0, such that, if $p \leq \kappa n$, then*

$$\sup_{\vartheta \in \Theta_{c_0}(n,p;\Lambda,\sigma)} \mathbb{E}_\vartheta \bar{h}(\widehat{\mathbf{z}}^{\mathrm{prof}}, \mathbf{z}) \leq C \min(1, r) + Ce^{-cn},$$

$$\sup_{\vartheta \in \Theta_{c_0}(n,p;\Lambda,\sigma)} \mathbb{P}_\vartheta(\widehat{\mathbf{z}}^{\mathrm{prof}} \notin \{\mathbf{z}, -\mathbf{z}\}) \leq Cnr + Ce^{-cn}.$$

Remark 2.2. The proof compares candidate partitions through the excess profile RSS. The signal term is lower bounded by a chi-square radius whose degrees of freedom are controlled by the Hamming error, while the fluctuation terms admit conditional exponential bounds in the same radius. Integrating over this radius and summing over Hamming shells yields the result. See Appendix B.

For the lower bound, consider the oracle experiment

$$(G, y), \qquad y = \frac{z}{2}\Lambda G + \sigma\xi, \qquad z \in \{+1,-1\}, \qquad G, \xi \overset{\mathrm{iid}}{\sim} N(0,1), \tag{1}$$

which is a symmetric subproblem of the original model with the regression vectors revealed. Conditional on G, recovering z is a binary Gaussian testing problem.

Lemma 2.3 (Oracle Bayes error)**.** *For the experiment* (1) *under the uniform prior on $z \in \{+1,-1\}$, the Bayes misclassification error is*

$$q(r) = \mathbb{E}[\Phi(-|G|/(2r))] = \frac{1}{\pi}\arctan(2r),$$

where Φ is the standard normal cdf, with the convention $q(0) = 0$.

Proposition 2.4 (Oracle lower bounds)**.** *There exists a constant $\zeta > 0$, depending only on c_0, such that, for all sufficiently large n,*

$$R_H^*(n,p,\Lambda,\sigma) \geq \frac{\lfloor \zeta n \rfloor}{n} q(r),$$

$$R_{\mathrm{ex}}^*(n,p,\Lambda,\sigma) \geq 1 - (1 - q(r))^{\lfloor \zeta n \rfloor}.$$

Remark 2.5. The proof embeds $\lfloor \zeta n \rfloor$ independent copies of the oracle experiment into a balanced subfamily. Under the product prior, the MAP error count follows a binomial distribution $\mathrm{Bin}(\lfloor \zeta n \rfloor, q(r))$. Accounting for global label interchange gives both lower bounds. See Appendix C.

Remark 2.6. Since $q(r) \asymp \min(1, r)$, Propositions 2.1 and 2.4 give the minimax mis-clustering scale in Theorem 1.1. Their exact-recovery bounds give the threshold $nr \to 0$. The two recovery scales reflect complementary aspects of the random-margin distribution. The order-r mis-clustering rate is governed by the mass of margins at the noise scale, whereas exact recovery depends on the smallest realized margin, with

$$n \min_{1 \leq i \leq n} \frac{|\mathbf{x}_i^\top \boldsymbol{\theta}|}{\Lambda} \xrightarrow{d} \mathrm{Exp}(\sqrt{2/\pi}),$$

where $\mathrm{Exp}(\lambda)$ is the exponential distribution with rate λ.

Figure 2 illustrates these two recovery scales for the algorithm in Section 3 under the reference setting of Section 6 and Appendix F. Across the displayed (n, p) pairs, mis-clustering error tracks $q(r)$, while exact-recovery probability is close to the curve $\exp(-2nr/\pi)$. For small r, this curve approximates the oracle probability $(1 - q(r))^n$ of recovering all memberships, since $q(r) \sim 2r/\pi$.

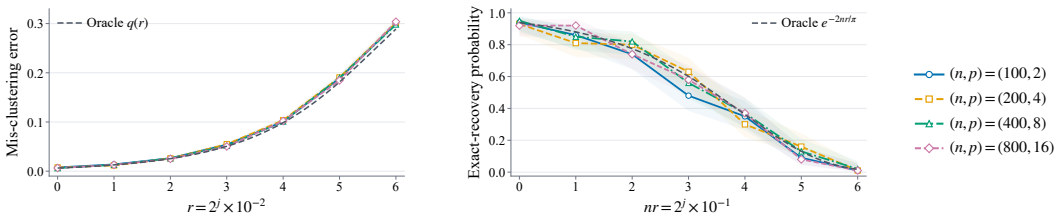


Figure 2: Recovery scaling across (n, p) pairs. Left: mis-clustering error versus r with $q(r)$. Right: exact recovery probability versus nr with $\exp(-2nr/\pi)$.

3 Recovery Algorithm

For $I \subseteq [n] := \{1, \dots, n\}$, let $\widehat{\boldsymbol{\beta}}(I) := \big(\sum_{i\in I} \mathbf{x}_i \mathbf{x}_i^\top\big)^{-1} \sum_{i\in I} \mathbf{x}_i y_i$, whenever the Gram matrix is invertible. For $T > 0$, define the bounded residual transform $h_T(u) := 1 - \exp(-u^2/(2T^2))$.

Algorithm 3.1 (Membership recovery)**.**

(i) **Spectral initialization.** Compute

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top, \qquad \widehat{\mathbf{b}} = \widehat{\boldsymbol{\beta}}([n]), \qquad \widehat{T}^2 = \frac{1}{n}\sum_{i=1}^{n} \big(y_i - \mathbf{x}_i^\top \widehat{\mathbf{b}}\big)^2.$$

Form the residual-weighted spectral matrix

$$\widehat{\mathbf{M}} = \frac{1}{n}\sum_{i=1}^{n} h_{\widehat{T}}\big(y_i - \mathbf{x}_i^\top \widehat{\mathbf{b}}\big)\big(\mathbf{x}_i \mathbf{x}_i^\top - \widehat{\boldsymbol{\Sigma}}\big).$$

Take a leading generalized eigenvector of $\widehat{\mathbf{M}}$,

$$\widehat{\mathbf{w}} \in \operatorname*{arg\,max}_{\mathbf{w}:\, \mathbf{w}^\top \widehat{\boldsymbol{\Sigma}} \mathbf{w} = 1} \mathbf{w}^\top \widehat{\mathbf{M}} \mathbf{w},$$

and initialize

$$z_i^{(0)} = \operatorname{sgn}\big((\mathbf{x}_i^\top \widehat{\mathbf{w}})(y_i - \mathbf{x}_i^\top \widehat{\mathbf{b}})\big), \qquad i \in [n].$$

(ii) **Hard least-squares refinement.** For $k = 0, \dots, K$, alternate between regression refitting and membership reassignment,

$$\boldsymbol{\gamma}_b^{(k)} = \widehat{\boldsymbol{\beta}}\big(\{i \in [n] : z_i^{(k)} = b\}\big), \qquad b \in \{+1, -1\},$$
$$z_i^{(k+1)} = \operatorname*{arg\,min}_{b\in\{+1,-1\}} \big|y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_b^{(k)}\big|, \qquad i \in [n].$$

Return $\widehat{\mathbf{z}} = \mathbf{z}^{(K+1)}$.

Remark 3.1. For $r > 0$, Theorem 1.2 requires $K \geq \lceil 5 \log(1/r) \rceil$. Algorithm 3.1 has computational complexity $O\big((K+1)(np^2 + p^3)\big)$. The RSS is nonincreasing under refinement; in practice, refinement may stop once the assignments no longer change. A rank-deficient least-squares refit is regarded as algorithmic failure.

The hard least-squares refinement follows the standard iterative approach of alternating groupwise least-squares refitting and residual-based membership reassignment (Yi et al., 2014; Bonhomme & Manresa, 2015). Section 4 establishes its local contraction, and Section 5 constructs an initializer satisfying the required local condition.

4 Local Contraction for Membership Recovery

For a coefficient pair $\gamma = (\boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-)$, define its normalized prediction error by

$$a(\gamma) := \frac{1}{\Lambda} \min_{\pi} \max_{b\in\{+,-\}} \big\|\boldsymbol{\gamma}_{\pi(b)} - \boldsymbol{\beta}_b^*\big\|_{\boldsymbol{\Sigma}}.$$

For $a(\gamma) < 1/2$, the minimizing permutation is unique, and throughout the local analysis we orient γ accordingly. Let $\widehat{z}_i(\gamma) \in \arg\min_{b\in\{+,-\}} |y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_b|$ denote the induced nearest-residual assignment, and define $H(\gamma) := n^{-1}\sum_{i=1}^n \mathbf{I}(\widehat{z}_i(\gamma) \neq z_i^*)$.

For the contraction argument, define the margin-weighted Hamming loss

$$H^{\mathrm{w}}(\gamma) := \frac{1}{n}\sum_{i=1}^n \frac{(\mathbf{x}_i^\top \boldsymbol{\theta})^2}{\Lambda^2}\mathbf{I}(\widehat{z}_i(\gamma) \neq z_i^*).$$

For the hard least-squares iterations, $\widehat{z}_i(\gamma^{(k)}) = z_i^{(k+1)}$.

Theorem 4.1 (Local margin-weighted contraction)**.** *There exist constants $C, \rho, r_0, \kappa > 0$, depending only on c_0, such that, for every $t \geq 1$ satisfying $p + t + \log(en) \leq \kappa n$ and $r \leq r_0$, with probability at least $1 - Ce^{-t}$, uniformly over all starting pairs $\gamma^{(0)}$ satisfying $a(\gamma^{(0)}) \leq \rho/2$, the hard least-squares iterates satisfy, for all $k \geq 0$,*

$$a(\gamma^{(k)}) \leq \rho,$$

and

$$H^{\mathrm{w}}(\gamma^{(k)}) \leq 2^{-k}H^{\mathrm{w}}(\gamma^{(0)}) + C\big(r^3 + r^{5/2}\sqrt{\frac{t}{n}} + r^2\frac{t}{n}\big).$$

Remark 4.2. The margin-weighted Hamming loss is motivated by Gao & Zhang (2022), where each label mismatch is weighted by the squared ℓ_2 distance between population signals. In our model, this yields the weight $(\mathbf{x}_i^\top \boldsymbol{\theta})^2/\Lambda^2$. Their conversion from weighted to ordinary Hamming loss uses a uniform lower bound on signal separation. Here, however, the separations are observation-specific random margins with nonnegligible mass near zero, so the same reduction would require control through the smallest realized margin, which is too crude for the sharp mis-clustering scale.

A key geometric observation is that, despite the absence of a uniform separation, misassignments induced by any local coefficient pair are confined to narrow Gaussian wedges around the margin direction; see Fact D.3 and Lemma D.6. Uniform control of these wedges keeps the assigned Gram matrices well conditioned and the iterates local, allowing the geometric bounds to control both the current and updated misassignment sets and yield the weighted-Hamming contraction in Lemma D.8.

The remaining small-margin and oracle-prediction contributions are captured by the iterate-independent remainder Q°, controlled in Lemma D.9. Together these bounds yield Theorem 4.1. Fact D.4 then converts the terminal weighted loss to ordinary Hamming error by separating off the empirical small-margin count, yielding the corollary below.

Corollary 4.3 (Local membership recovery)**.** *Let ρ, r_0, κ be as in Theorem 4.1. Suppose $0 < r \leq r_0$, $p + \log(en) \leq \kappa n$, and let $K \geq \lceil 5\log(1/r)\rceil$. Then, uniformly over all starting pairs $\gamma^{(0)}$ satisfying $a(\gamma^{(0)}) \leq \rho/2$,*

$$\mathbb{E}[H(\gamma^{(K)})] \leq Cr + Ce^{-cn},$$
$$\mathbb{P}\big(H(\gamma^{(K)}) > 0\big) \leq Cnr + Ce^{-cn}.$$

Moreover, uniformly over the same starting pairs, for every $t \geq 1$ satisfying $p + t + \log(en) \leq \kappa n$,

$$\mathbb{P}\Big(H(\gamma^{(K)}) > C\Psi_{n,r}(t)\Big) \leq Ce^{-t}.$$

Remark 4.4. Combined with the initialization guarantee in Section 5, Corollary 4.3 gives the finite-sample Hamming tail scale $\Psi_{n,r}(t)$ for the estimator. In the regime $nr \gtrsim 1$ with small $r > 0$, this scale is minimax sharp up to constants; see Corollary D.10. When $nr \lesssim 1$, the tail at level $1/n$ corresponds to exact-recovery failure, whose optimal probability is of order nr.

The proofs of Theorem 4.1 and Corollary 4.3 are given in Appendix D.

5 Initialization

The local contraction result reduces efficient recovery to constructing an initial partition whose first hard least-squares refit enters the required neighborhood. We motivate such an initialization from the noiseless population setting.

Population sign rule. Let $\omega := |I_+^*|/n \in [c_0, 1-c_0]$ and define the population pooled least-squares coefficient $\bar{\boldsymbol{\beta}} := \omega\boldsymbol{\beta}_+^* + (1-\omega)\boldsymbol{\beta}_-^*$. The corresponding residual $R := y - \mathbf{x}^\top\bar{\boldsymbol{\beta}}$ satisfies

$$R = \begin{cases} (1-\omega)\mathbf{x}^\top\boldsymbol{\theta} + \epsilon, & z^* = +, \\ -\omega\mathbf{x}^\top\boldsymbol{\theta} + \epsilon, & z^* = -. \end{cases}$$

When $\epsilon = 0$, $\operatorname{sgn}((\mathbf{x}^\top\boldsymbol{\theta})R) = z^*$ almost surely. Thus an accurate pooled coefficient and separation direction naturally yield an initial partition through the same sign rule. The pooled least-squares estimator $\widehat{\mathbf{b}}$ concentrates around $\bar{\boldsymbol{\beta}}$ by Lemma E.2, so it remains to estimate the direction of $\boldsymbol{\theta}$.

Spectral identification of the separation direction. We estimate the separation direction from the pooled residuals $y_i - \mathbf{x}_i^\top\widehat{\mathbf{b}}$. At the population pooled coefficient, the corresponding residual R satisfies $\mathbb{E}[\mathbf{x}R] = 0$, so we instead consider a second-order residual moment. If $\boldsymbol{\Sigma} = \mathbf{I}_p$, then

$$\begin{aligned}\mathbb{E}\big[R^2(\mathbf{x}\mathbf{x}^\top - \mathbf{I}_p)\big] &= \big(\omega(1-\omega)^2 + (1-\omega)\omega^2\big)\mathbb{E}\big[(\mathbf{x}^\top\boldsymbol{\theta})^2(\mathbf{x}\mathbf{x}^\top - \mathbf{I}_p)\big] \\ &= 2\omega(1-\omega)\boldsymbol{\theta}\boldsymbol{\theta}^\top,\end{aligned}$$

by the Gaussian fourth-moment identity. For general $\boldsymbol{\Sigma}$,

$$\mathbb{E}\big[R^2(\mathbf{x}\mathbf{x}^\top - \boldsymbol{\Sigma})\big] = 2\omega(1-\omega)(\boldsymbol{\Sigma}\boldsymbol{\theta})(\boldsymbol{\Sigma}\boldsymbol{\theta})^\top.$$

Since $\boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\theta} = \Lambda^2$, $\boldsymbol{\theta}$ is the leading generalized eigenvector relative to $\boldsymbol{\Sigma}$.

Remark 5.1. The construction is related to spectral and moment methods for mixed linear regression (Chaganty & Liang, 2013; Yi et al., 2014). A key distinction is that our initializer targets only the separation direction rather than the two regression components. The rank-one population second moment directly identifies $\boldsymbol{\theta}$, avoiding a separate component-identification step. The individual regression coefficients are then estimated by the subsequent hard least-squares refit.

Bounded residual weighting. Squared-residual weighting introduces a logarithmic loss in concentration of the empirical spectral matrix; see Proposition E.3. This motivates the use of the bounded transform h_T in Algorithm 3.1. Facts E.5–E.6 show that, at the population pooled coefficient, bounded weighting preserves the rank-one structure and identifies $\boldsymbol{\theta}$.

Remark 5.2. The bounded transform provides the uniform spectral control needed when p is of the same order as n in Lemma E.9. Empirically, it improves the spectral direction estimation as p/n increases, although the two weighting schemes perform similarly after hard least-squares refinement except at the largest p/n; see Figure 7 in Appendix F.

Theorem 5.3. *Let ρ be as in Theorem 4.1. There exist constants $C, \delta_0, r_0, \kappa > 0$, depending only on c_0, such that for every $t \geq 1$ satisfying $p + t + \log(en) \leq \kappa n$ and $r \leq r_0$, with probability at least $1 - Ce^{-t}$, the spectral direction satisfies*

$$\sin\angle(\boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{w}}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}) \leq \delta_0,$$

where $\angle(\cdot,\cdot)$ denotes the sign-invariant angle. Moreover, on the same event, the first least-squares refit in Algorithm 3.1 satisfies

$$a(\gamma^{(0)}) \leq \rho/2.$$

Remark 5.4. For initialization, it suffices to estimate the separation direction within a fixed angle. Together with an accurate pooled coefficient, this confines sign-rule misassignments to the Gaussian wedges analyzed in Section 4, and Lemma E.10 then shows that the first hard least-squares refit satisfies the required local-neighborhood condition.

Appendix E proves Theorem 5.3 and completes the proofs of Theorem 1.2 and Corollary 1.3.

6 Numerical experiments

We focus on a reference Gaussian setting, with additional simulations reported in Appendix F.

Simulation setting. Unless stated otherwise, the baseline setting uses balanced groups with $n = 400$, $p = 8$, and $\Lambda = 1$. We set $\boldsymbol{\beta}_\pm^* = \boldsymbol{\mu} \pm \boldsymbol{\theta}/2$ with $\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} = 2\Lambda$ and $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}\boldsymbol{\theta} = 0$. This construction avoids the special symmetric case $\boldsymbol{\beta}_+^* = -\boldsymbol{\beta}_-^*$ while ensuring that the two regression vectors have equal prediction norms. The design covariance is set with condition number $\kappa_{\boldsymbol{\Sigma}} = 5$. Results are based on

100 Monte Carlo replications and are summarized by the permutation-adjusted mis-clustering error and exact-recovery probability.

We compare three classes of methods. For hard least-squares refinement, we compare our partition initializer with random starts, which selects the partition with the smallest RSS, and SVD-grid (Yi et al., 2014), adapted to general covariance to initialize the regression components. For clusterwise least squares, we use multi-start Späth (Späth, 1979) and GA–Lloyd (Park et al., 2017). We also compare the subgroup identification method of Ma et al. (2020), which fuses observation-specific coefficients through a pairwise concave penalty, with either its original or the proposed initialization.

The random-start initializer uses 100 initial partitions, whereas multi-start Späth uses 10 starts, as the exchange search is more computationally intensive. Pointwise 95% Monte Carlo confidence intervals use a normal approximation for mean mis-clustering error and Wilson intervals for exact-recovery probability. We also evaluate Spectral Experts (Chaganty & Liang, 2013) and the convex method of Chen et al. (2014) as coefficient initializers for hard least-squares refinement. Their results, together with further implementation details, are deferred to Appendix F, since neither improves on the leading methods in the reference setting.

Noiseless recovery. For $r = 0$, Figure 3 varies p and n. The proposed method achieves exact recovery over the widest range of both quantities. As p approaches the within-group sample size, groupwise least-squares fits become unstable and recovery deteriorates across methods.

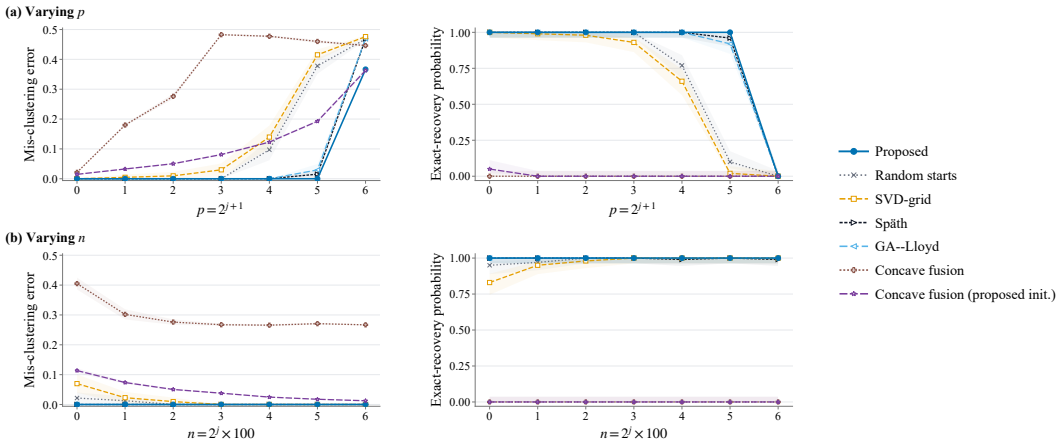


Figure 3: Noiseless recovery with varying p and n.

Noisy recovery. For $r > 0$, Figure 4 reports four experiments under the reference setting. The first two vary the theoretical scales r and nr, while the remaining two vary p and n at fixed $r = 0.1$.

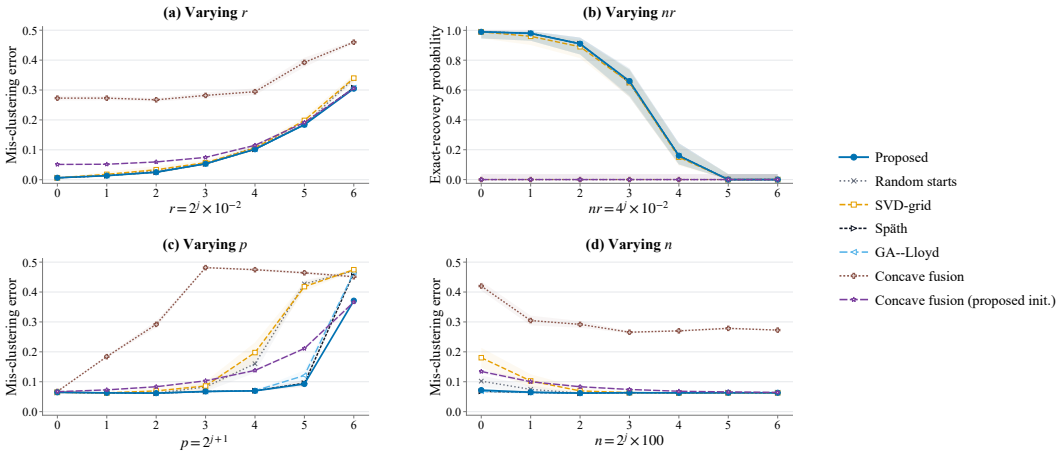


Figure 4: Noisy recovery under Gaussian noise with varying r, nr, p, and n.

The first two panels show the mis-clustering and exact-recovery transitions on the r and nr scales, respectively. As p approaches the within-group sample size, recovery again deteriorates because of unstable groupwise least-squares fits. At fixed $r = 0.1$, increasing n improves SVD-grid, random starts, and concave fusion with the proposed initialization. Even at larger n, the strongest methods have nonzero mis-clustering error, reflecting the intrinsic recovery limit under fixed nonzero noise from Theorem 1.1.

Across both settings, among methods using the hard least-squares refinement, the proposed initialization outperforms SVD-grid and random starts. It also performs as well as or better than multi-start Späth and GA–Lloyd. The proposed initialization substantially improves concave fusion, but it still fails to achieve exact recovery.

Table 1: Mean running time in seconds at three representative settings.

	$(n,p) = (400, 8)$		$(n,p) = (400, 128)$		$(n,p) = (6400, 8)$	
Method	$r = 0$	$r = 0.1$	$r = 0$	$r = 0.1$	$r = 0$	$r = 0.1$
Proposed	**0.0008**	**0.0010**	**0.0442**	**0.0488**	**0.0034**	**0.0044**
Random starts	0.0084	0.0094	0.0915	0.1006	0.0270	0.0304
SVD-grid	0.0018	0.0021	0.0485	0.0535	0.0101	0.0111
Späth	0.7265	0.7780	5.9383	5.9347	10.7358	13.6481
GA–Lloyd	0.1034	0.1127	6.2251	5.8839	0.5470	0.7158
Concave fusion	0.1215	0.1760	0.1930	0.2267	41.7158	41.1869
Concave fusion (proposed init.)	0.0044	0.0211	0.2878	0.2937	1.0757	1.0629

With two groups fixed in our experiments, we report the running time for concave fusion at the selected tuning parameter rather than over the full grid. Appendix F.2 reports the complete timing results, where the proposed method is the fastest across all settings considered. Table 1 shows three representative settings. Multi-start Späth, GA–Lloyd, and concave fusion are substantially more expensive, particularly as n or p increases. Appendix F also provides full implementation details and additional robustness experiments.

7 Discussion

Our results show that, in the two-component Gaussian random-design model studied here, the sharp limits of latent membership recovery are governed by the random regression margins, and that these limits can be achieved efficiently. The same minimax mis-clustering scale and exact-recovery threshold also extend to random memberships. If the labels are drawn independently with group probabilities bounded away from zero and one, the realized groups are balanced with high probability, and the fixed-membership results apply conditional on this event. More generally, the oracle problem in Section 2 suggests that the recovery scales are determined by the small-ball behavior of the random margin. If $\mathbb{P}(|G| \le t) \asymp t^\alpha$ near zero, then under Gaussian noise the corresponding scales become r^α for mis-clustering and nr^α for exact recovery; the $\alpha = 2$ experiment in Appendix F illustrates this behavior.

For a fixed number $K_0 \ge 3$ of latent groups with balanced sizes, we believe the same order-r mis-clustering scale and exact-recovery threshold $nr \to 0$ continue to hold, with r defined by the smallest $\mathbf{\Sigma}$-norm separation between any two regression components. The oracle lower bound, profile-RSS upper bound, and local hard least-squares analysis can be extended with essentially the same arguments. For initialization, the population spectral structure identifies a centered coefficient-contrast subspace of dimension at most $K_0 - 1$, and it remains unclear how to construct a suitable initializer from this subspace without imposing additional structural conditions.

A broader implication concerns statistical inference after group recovery. Many subgroup and latent-group methods use estimated memberships for subsequent estimation and inference (Ma & Huang, 2017; Ma et al., 2020; Bonhomme & Manresa, 2015; Su et al., 2016). Our results suggest that exact recovery may be stronger than necessary for subsequent inference in cross-sectional mixed regression, since membership errors can persist under nonvanishing noise even when the regression components are accurately estimated. This motivates inference procedures that account for membership uncertainty. Repeated measurements or multiple data sources may improve recovery by providing more information for each membership. Inference for the memberships themselves, such as confidence sets, is another important direction (Dzemski & Okui, 2024).

Reproducibility Statement

Complete proofs and supporting technical results are provided in Appendices A–E. Section 6 and Appendix F describe the simulation designs, comparison methods, and implementation details. Code for reproducing all reported simulation figures and tables is available in an anonymous repository.

AI Use Statement

Generative AI tools were used in this work to assist with literature search, language editing and polishing, and preliminary drafting of code. All mathematical results and arguments were developed and verified by the authors. Any AI-assisted code was independently reviewed, tested, corrected, and adapted by the authors before being used in this work. The authors take full responsibility for the content of this work.

References

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. doi: 10.1214/009053606000001217.

Martin Azizyan, Aarti Singh, and Larry A. Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2139–2147, 2013.

Adarsh Barik and Jean Honorio. Sparse mixed linear regression with guarantees: Taming an intractable problem with invex relaxation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1627–1646, 2022.

Peter L. Bartlett and Gábor Lugosi. An inequality for uniform deviations of sample averages from their means. *Statistics & Probability Letters*, 44(1):55–62, 1999. doi: 10.1016/S0167-7152(98)00291-0.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015. doi: 10.3982/ECTA11319.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1040–1048, 2013.

Xiaohui Chen and Yun Yang. Cutoff for exact recovery of gaussian mixture models. *IEEE Transactions on Information Theory*, 67(6):4223–4238, 2021. doi: 10.1109/TIT.2021.3063155.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 560–604, 2014.

Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5): 822–883, 2017. doi: 10.1002/cpa.21638.

Richard D. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8 (3):227–245, 1989. doi: 10.1016/0167-9473(89)90043-1.

Wayne S. DeSarbo and William L. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282, 1988. doi: 10.1007/BF01897167.

Luc Devroye, Silvio Lattanzi, Gábor Lugosi, and Nikita Zhivotovskiy. On mean estimation for heteroscedastic random variables. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 59(1):1–20, 2023. doi: 10.1214/21-AIHP1239.

Andreas Dzemski and Ryo Okui. Confidence set for group membership. *Quantitative Economics*, 15 (2):245–277, 2024. doi: 10.3982/QE1975.

Yingjie Fei and Yudong Chen. Hidden integrality of SDP relaxations for sub-gaussian mixture models. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1931–1965, 2018.

Chao Gao and Anderson Y. Zhang. Iterative algorithm for discrete structure recovery. *The Annals of Statistics*, 50(2):1066–1094, 2022. doi: 10.1214/21-AOS2140.

Avishek Ghosh and Kannan Ramchandran. Alternating minimization converges super-linearly for mixed linear regression. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1093–1103, 2020.

Paul Hand and Babhru Joshi. A convex program for mixed linear regression with a recovery guarantee for well-separated data. *Information and Inference: A Journal of the IMA*, 7(3):563–579, 2018. doi: 10.1093/imaiai/iax018.

Christian Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296, 2000. doi: 10.1007/s003570000022.

Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1405–1413, 2021.

Matthias Löffler, Anderson Y. Zhang, and Harrison H. Zhou. Optimality of spectral clustering in the gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, 2021. doi: 10.1214/20-AOS2044.

Yu Lu and Harrison H. Zhou. Statistical and computational guarantees of Lloyd's algorithm and its variants, 2016. arXiv:1612.02099.

Shujie Ma and Jian Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017. doi: 10.1080/01621459.2016.1148039.

Shujie Ma, Jian Huang, Zhiwei Zhang, and Mingming Liu. Exploration of heterogeneous treatment effects via concave fusion. *The International Journal of Biostatistics*, 16(1):20180026, 2020. doi: 10.1515/ijb-2018-0026.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999. doi: 10.1214/aos/1017939240.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5):2326–2366, 2006. doi: 10.1214/009053606000000786.

Mohamed Ndaoud. Sharp optimal recovery in the two component gaussian mixture model. *The Annals of Statistics*, 50(4):2096–2126, 2022. doi: 10.1214/22-AOS2178.

Young Woong Park, Yan Jiang, Diego Klabjan, and Loren Williams. Algorithms for generalized clusterwise linear regression. *INFORMS Journal on Computing*, 29(2):301–317, 2017. doi: 10.1287/ijoc.2016.0729.

Richard E. Quandt and James B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):730–738, 1978. doi: 10.1080/01621459.1978.10480085.

Helmuth Späth. Algorithm 39: Clusterwise linear regression. *Computing*, 22(4):367–373, 1979. doi: 10.1007/BF02265317.

Liangjun Su, Zhentao Shi, and Peter C. B. Phillips. Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264, 2016. doi: 10.3982/ECTA12560.

Nelvin Tan and Ramji Venkataramanan. Mixed linear regression via approximate message passing. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2116–2131, 2023.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. doi: 10.1214/aos/1079120131.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Jon A. Wellner. The Bennett-Orlicz norm. *Sankhya A*, 79(2):355–383, 2017. doi: 10.1007/s13171-017-0108-4.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 613–621, 2014.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008.

Kai Zhong, Prateek Jain, and Inderjit S. Dhillon. Mixed linear regression with multiple components. In *Advances in Neural Information Processing Systems 29*, pp. 2190–2198, 2016.

The appendices are organized as follows. Appendix A introduces notation and auxiliary results used across proofs. Appendices B and C prove the information-theoretic upper and lower bounds. Appendix D proves the local contraction and membership-recovery results, and Appendix E analyzes the initialization and completes the efficient-recovery proof. Appendix F gives implementation details and additional numerical results.

Throughout the appendices, elementary identities and deterministic bounds are stated as Facts before they are used in the proofs and, when shared across several arguments, may be collected at the beginning of an appendix.

A Notation and auxiliary results

A.1 Notation

For $b \in \{+,-\}$, write $I_b^* := \{i \in [n] : z_i^* = b\}$ for the two true groups. For two label vectors $\mathbf{s}, \mathbf{z}$ of the same length, denote their Hamming distance by $d_H(\mathbf{s},\mathbf{z}) := \sum_i \mathbf{I}(s_i \neq z_i)$. For a vector $\mathbf{h}$, $\|\mathbf{h}\|$ denotes its Euclidean norm; for a matrix $\mathbf{H}$, $\|\mathbf{H}\|_{\mathrm{op}}$ and $\|\mathbf{H}\|_{\mathrm{F}}$ denote its operator and Frobenius norms. For nonnegative quantities, $\lesssim$ and $\gtrsim$ denote inequalities up to a constant factor, and $\asymp$ denotes equivalence up to constant factors. Throughout the appendices, $C, c > 0$ denote generic constants whose values may change from line to line.

Write $\mathbf{y} := (y_i)_{i=1}^n$ and $\boldsymbol{\epsilon} := (\epsilon_i)_{i=1}^n$. When an observation-level random variable is written without an index, it denotes a generic copy with the same distribution as the corresponding indexed variable.

For nonzero vectors $\mathbf{h}, \mathbf{q}$, define the sign-invariant angle by

$$\angle(\mathbf{h},\mathbf{q}) := \arccos \frac{|\mathbf{h}^\top \mathbf{q}|}{\|\mathbf{h}\|\,\|\mathbf{q}\|}.$$

For two subspaces $\mathcal{S}, \mathcal{T}$ of the same dimension, let $\Theta(\mathcal{S},\mathcal{T})$ denote their principal angles. If $\mathbf{Q}_{\mathcal{S}}$ has orthonormal columns spanning $\mathcal{S}$ and $\mathbf{P}_{\mathcal{T}}$ is the orthogonal projector onto $\mathcal{T}$, then

$$\|\sin\Theta(\mathcal{S},\mathcal{T})\|_{\mathrm{op}} = \|(\mathbf{I}-\mathbf{P}_{\mathcal{T}})\mathbf{Q}_{\mathcal{S}}\|_{\mathrm{op}}, \qquad \|\sin\Theta(\mathcal{S},\mathcal{T})\|_{\mathrm{F}} = \|(\mathbf{I}-\mathbf{P}_{\mathcal{T}})\mathbf{Q}_{\mathcal{S}}\|_{\mathrm{F}}.$$

Whitened coordinates. Throughout the appendices, we use the whitening transformation $\mathbf{A} := \boldsymbol{\Sigma}^{1/2}$, $\mathbf{u}_i := \mathbf{A}^{-1}\mathbf{x}_i$, and $\boldsymbol{\alpha} := \mathbf{A}\boldsymbol{\beta}$. Let $\mathbf{X}$ and $\mathbf{U}$ denote the design matrices with rows $\mathbf{x}_i^\top$ and $\mathbf{u}_i^\top$, respectively, so that $\mathbf{X} = \mathbf{U}\mathbf{A}$. Then

$$\mathbf{u}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{I}_p), \qquad \mathbf{x}_i^\top\boldsymbol{\beta} = \mathbf{u}_i^\top\boldsymbol{\alpha}, \qquad \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_{\boldsymbol{\Sigma}} = \|\mathbf{A}(\boldsymbol{\beta}-\boldsymbol{\beta}')\|.$$

For $I \subset [n]$, let $\mathbf{X}_I$ and $\mathbf{U}_I$ denote the corresponding row-restricted matrices. Then $\mathbf{X}_I = \mathbf{U}_I\mathbf{A}$.

Since $\boldsymbol{\beta} \mapsto \boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\beta}$ is bijective and $\mathbf{x}_i^\top\boldsymbol{\beta} = \mathbf{u}_i^\top\boldsymbol{\alpha}$, the least-squares objectives, fitted values such as $\mathrm{RSS}(\mathbf{z})$ and nearest-residual assignments are identical in the two coordinate systems.

For the true regression coefficients, write $\boldsymbol{\alpha}_b^* := \mathbf{A}\boldsymbol{\beta}_b^*$ for $b \in \{+,-\}$, and define $\mathbf{v} := \mathbf{A}\boldsymbol{\theta}/\Lambda$. Since $\boldsymbol{\theta} = \boldsymbol{\beta}_+^* - \boldsymbol{\beta}_-^*$ and $\Lambda = \|\boldsymbol{\theta}\|_{\boldsymbol{\Sigma}}$, we have $\boldsymbol{\alpha}_+^* - \boldsymbol{\alpha}_-^* = \Lambda\mathbf{v}$, with $\|\mathbf{v}\| = 1$.

For the population pooled coefficient from Section 5, write $\bar{\boldsymbol{\alpha}} := \mathbf{A}\bar{\boldsymbol{\beta}}$.

Write $\epsilon_i = \Lambda r \xi_i$, where $\xi_i \overset{\text{iid}}{\sim} N(0,1)$ independently of the design. When $r = 0$, the ξ_i are taken to be auxiliary iid standard Gaussian variables independent of the design.

Finally, define the full-sample whitened Gram matrix $\widehat{\mathbf{S}} := n^{-1}\mathbf{U}^\top\mathbf{U}$. We have $\widehat{\boldsymbol{\Sigma}} = \mathbf{A}\widehat{\mathbf{S}}\mathbf{A}$.

A.2 Uniform VC inequalities

Let $W_1, \ldots, W_n$ be iid random elements with distribution P. For a class $\mathcal{A}$ of measurable sets, define

$$P_n(A) := \frac{1}{n}\sum_{i=1}^n \mathbf{I}(W_i \in A), \qquad \psi_n(V,t) := \frac{V\log(en/V)+t}{n}.$$

Any Boolean combination involving a fixed number of affine halfspaces in $\mathbb{R}^d$ has VC dimension $O(d)$ (see van der Vaart & Wellner, 1996, Lemmas 2.6.17–2.6.18).

We use the following standard absolute and relative VC inequalities. Relative-deviation VC bounds originate with Vapnik & Chervonenkis (1971); see also Bartlett & Lugosi (1999).

Lemma A.1. *Suppose $\mathcal{A}$ is pointwise measurable and has VC dimension at most V, where $1 \le V \le n$. Then, for every $t > 0$, with probability at least $1 - e^{-t}$, the following hold simultaneously:*

(i)

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \le C\sqrt{\psi_n(V,t)}.$$

(ii) For every $A \in \mathcal{A}$,

$$P_n(A) \le P(A) + C\sqrt{P(A)\psi_n(V,t)} + C\psi_n(V,t).$$

Proof. Part (i) is the standard VC inequality. For part (ii), applying Lemma 2 of Devroye et al. (2023) to the VC class $\{\mathbf{I}(A) : A \in \mathcal{A}\}$ with $\delta = e^{-t}$ gives, simultaneously for all $A \in \mathcal{A}$,

$$|P_n(A) - P(A)| \le C\sqrt{P(A)\psi_n(V,t)} + C\psi_n(V,t),$$

which implies the stated bound. □

A.3 Uniform Gaussian-design bounds

Lemma A.2. *For every $t > 0$, with probability at least $1 - Ce^{-t}$, the following hold simultaneously,*

(i)

$$\left\|n^{-1}\mathbf{U}^\top\mathbf{U} - \mathbf{I}_p\right\|_{\mathrm{op}} \le C\big(\sqrt{(p+t)/n} + (p+t)/n\big).$$

In particular, if $p + t \le \kappa n$ for sufficiently small κ, then the eigenvalues of $n^{-1}\mathbf{U}^\top\mathbf{U}$ are bounded away from zero and infinity by fixed constants.

(ii) Simultaneously for every nonempty $I \subset [n]$,

$$\frac{1}{n}\|\mathbf{U}_I\|_{\mathrm{op}}^2 \le C\left(\frac{p+t}{n} + \frac{|I|}{n}\log\frac{en}{|I|}\right).$$

Proof. Part (i) follows from Theorem 4.6.1 of Vershynin (2018). For part (ii), fix $m \in [n]$ and a set $I \subset [n]$ with $|I| = m$. The same theorem gives, for every $x > 0$,

$$\mathbb{P}(\|\mathbf{U}_I\|_{\mathrm{op}}^2 > C(p + m + x)) \le Ce^{-x}.$$

Take $x_m = t + 2m\log(en/m)$. Since $\binom{n}{m} \le (en/m)^m$, a union bound over all subsets of size m gives failure probability at most

$$C\binom{n}{m}e^{-x_m} \le Ce^{-t}\exp\big(-m\log(en/m)\big) \le Ce^{-t-m}.$$

Summing over $m \in [n]$ gives a failure probability at most Ce^{-t}. Hence, simultaneously for every nonempty $I \subset [n]$,

$$\|\mathbf{U}_I\|_{\mathrm{op}}^2 \le C\left(p + t + |I|\log\frac{en}{|I|}\right),$$

since $|I| \le |I|\log(en/|I|)$. Intersecting this event with part (i) gives the stated probability bound after enlarging C. □

Lemma A.3 (Uniform conditioning of large subsets)**.** *Fix $\tau \in (0,1)$. There exist constants $c, C, \kappa > 0$, depending only on τ, such that, if $p \le \kappa n$, then with probability at least $1 - Ce^{-cn}$,*

$$\lambda_{\min}\left(\frac{1}{n}\mathbf{U}_E^\top\mathbf{U}_E\right) \ge c,$$

simultaneously for every $E \subset [n]$ with $|E| \ge \tau n$.

Proof. Choose $t_0 > 0$ such that $\mathbb{P}(|G| \le t_0) < \tau/4$ for $G \sim N(0,1)$. The class of centered slabs $\{\mathbf{u} : |\mathbf{q}^\top \mathbf{u}| \le t_0,\ \|\mathbf{q}\| = 1\}$ has VC dimension $O(p)$. If $p \le \kappa n$, then for $t = cn$,

$$\psi_n(Cp, cn) \le C\kappa \log \frac{e}{\kappa} + c.$$

Lemma A.1(i) therefore gives, with probability at least $1 - e^{-cn}$,

$$\sup_{\|\mathbf{q}\|=1} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(|\mathbf{q}^\top \mathbf{u}_i| \le t_0) - \mathbb{P}(|G| \le t_0) \right| \le C\sqrt{\kappa \log(e/\kappa) + c}.$$

For sufficiently small $c, \kappa > 0$, depending only on τ, the right-hand side is smaller than $\tau/4$. Hence

$$\sup_{\|\mathbf{q}\|=1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(|\mathbf{q}^\top \mathbf{u}_i| \le t_0) < \frac{\tau}{2}.$$

On this event, for every unit $\mathbf{q}$ and every $E \subset [n]$ with $|E| \ge \tau n$, at least $\tau n/2$ indices $i \in E$ satisfy $|\mathbf{q}^\top \mathbf{u}_i| > t_0$. Therefore

$$\mathbf{q}^\top \mathbf{U}_E^\top \mathbf{U}_E \mathbf{q} \ge \frac{\tau t_0^2}{2} n.$$

Taking the infimum over unit $\mathbf{q}$ and decreasing c if necessary so that $c \le \tau t_0^2/2$ proves the claim. □

Corollary A.4 (Gaussian sample bounds)**.** *There exist constants $C, c, \kappa > 0$, depending only on c_0, such that, for every $t \ge 1$ satisfying $p + t + \log(en) \le \kappa n$, there is an event $\mathcal{E}_{\mathrm{G}}(t)$ of probability at least $1 - Ce^{-t}$, on which the following hold simultaneously,*

$$\|\widehat{\mathbf{S}} - \mathbf{I}_p\|_{\mathrm{op}} \le C\sqrt{\frac{p+t}{n}}, \tag{2}$$

$$c\mathbf{I}_p \preceq \frac{1}{n} \mathbf{U}_{I_b^*}^\top \mathbf{U}_{I_b^*} \preceq C\mathbf{I}_p,\ b \in \{+,-\}, \tag{3}$$

$$c\mathbf{I}_p \preceq \frac{1}{n} \mathbf{U}_{I_b^* \setminus \{i\}}^\top \mathbf{U}_{I_b^* \setminus \{i\}} \preceq C\mathbf{I}_p,\ b \in \{+,-\},\ i \in I_b^*, \tag{4}$$

$$\max_{1 \le i \le n} \|\mathbf{u}_i\|^2 \le C(p + t + \log(en)), \tag{5}$$

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i^2 \le C, \quad \max_{b \in \{+,-\}} \Big\| \frac{1}{n} \sum_{i \in I_b^*} \mathbf{u}_i \xi_i \Big\| \le C\sqrt{\frac{p+t}{n}}. \tag{6}$$

Proof. (2) and (3) are direct applications of Lemma A.2(i). For (4), apply the same bound to every leave-one-out true-group design with confidence parameter $t + \log(en)$ and take a union bound.

Lemma A.2(ii), applied to the singleton subsets $I = \{i\}$, gives (5). Finally, applying Lemma A.2(i) to the augmented Gaussian designs with rows $(\mathbf{u}_i^\top, \xi_i)$ gives (6).

□

B PROFILE UPPER BOUND

This section proves Proposition 2.1. We use the whitening notation of Appendix A throughout.

For a candidate partition $\mathbf{z} \in \mathcal{T}_{c_0/2}$, the two groupwise least-squares problems can be combined into a single problem with block design matrix

$$\mathbf{D}_{\mathbf{z}} := \big[\mathrm{diag}\big(\mathbf{I}(z_i = +)\big)_{i=1}^n \mathbf{U} \ \big|\ \mathrm{diag}\big(\mathbf{I}(z_i = -)\big)_{i=1}^n \mathbf{U} \big].$$

Let $\mathbf{P}_{\mathbf{z}}$ be the orthogonal projector onto $\mathrm{col}(\mathbf{D}_{\mathbf{z}})$, and write $\mathbf{D}_* := \mathbf{D}_{\mathbf{z}^*}$ and $\mathbf{P}_* := \mathbf{P}_{\mathbf{z}^*}$ for the true partition. We first give two elementary facts that will be used in the proof of Proposition 2.1.

Fact B.1 (Profile RSS identities)**.** *For every $\mathbf{z} \in \mathcal{T}_{c_0/2}$,*

$$\mathrm{RSS}(\mathbf{z}) = \|(\mathbf{I} - \mathbf{P}_{\mathbf{z}})\mathbf{y}\|^2.$$

Writing $\mathbf{y} = \boldsymbol{\nu} + \boldsymbol{\epsilon}$, where $\nu_i = \mathbf{u}_i^\top \boldsymbol{\alpha}_{z_i^*}^*$, and setting $d_{\mathbf{z}}^2 := \|(\mathbf{I} - \mathbf{P}_{\mathbf{z}})\boldsymbol{\nu}\|^2$, we have

$$\mathrm{RSS}(\mathbf{z}) - \mathrm{RSS}(\mathbf{z}^*) = d_{\mathbf{z}}^2 + 2\boldsymbol{\epsilon}^\top (\mathbf{I} - \mathbf{P}_{\mathbf{z}})\boldsymbol{\nu} + \boldsymbol{\epsilon}^\top (\mathbf{P}_* - \mathbf{P}_{\mathbf{z}})\boldsymbol{\epsilon}. \tag{7}$$

Proof. The first identity follows directly from the least-squares formulation with design matrix $\mathbf{D}_\mathbf{z}$. Since $\boldsymbol{\nu} \in \mathrm{col}(\mathbf{D}_*)$, we have $\mathbf{P}_*\boldsymbol{\nu} = \boldsymbol{\nu}$ and hence $\mathrm{RSS}(\mathbf{z}^*) = \boldsymbol{\epsilon}^\top(\mathbf{I} - \mathbf{P}_*)\boldsymbol{\epsilon}$. Using $\mathbf{y} = \boldsymbol{\nu} + \boldsymbol{\epsilon}$ and the idempotence of $\mathbf{I} - \mathbf{P}_\mathbf{z}$ then gives (7). □

Fact B.2 (Quadratic minimization). *Let $\mathbf{H}, \mathbf{G} \succeq 0$ with $\mathbf{H} + \mathbf{G}$ invertible. Then, for every $\mathbf{d} \in \mathbb{R}^p$,*

$$\min_{\mathbf{a}\in\mathbb{R}^p} \big(\mathbf{a}^\top\mathbf{H}\mathbf{a} + (\mathbf{a}-\mathbf{d})^\top\mathbf{G}(\mathbf{a}-\mathbf{d})\big) = \mathbf{d}^\top\big[\mathbf{G} - \mathbf{G}(\mathbf{H}+\mathbf{G})^{-1}\mathbf{G}\big]\mathbf{d}.$$

Moreover, if $\mathbf{H} \succeq c\mathbf{G}$ for some $c > 0$, then

$$\mathbf{G} - \mathbf{G}(\mathbf{H}+\mathbf{G})^{-1}\mathbf{G} \succeq \frac{c}{1+c}\mathbf{G}.$$

Proof. The first identity follows by solving the normal equation $(\mathbf{H}+\mathbf{G})\mathbf{a} = \mathbf{G}\mathbf{d}$ and substituting the resulting minimizer. If $\mathbf{H} \succeq c\mathbf{G}$, then

$$\mathbf{a}^\top\mathbf{H}\mathbf{a} + (\mathbf{a}-\mathbf{d})^\top\mathbf{G}(\mathbf{a}-\mathbf{d}) \geq c\mathbf{a}^\top\mathbf{G}\mathbf{a} + (\mathbf{a}-\mathbf{d})^\top\mathbf{G}(\mathbf{a}-\mathbf{d}) \geq \frac{c}{1+c}\mathbf{d}^\top\mathbf{G}\mathbf{d}.$$

Minimizing over $\mathbf{a}$ and using the first identity gives the second inequality. □

Throughout the proof, for each candidate, choose the representative in $\{\mathbf{z}, -\mathbf{z}\}$ closest to $\mathbf{z}^*$ and write $m := d_H(\mathbf{z}, \mathbf{z}^*) \leq n/2$.

Proof idea. By (7) in Fact B.1, the excess RSS over the true partition decomposes into a signal term and two Gaussian fluctuation terms. On a uniform design event, the signal term is bounded below by a chi-square radius determined by the Hamming error, while the fluctuation terms admit conditional exponential tails in the same radius. The following lemma averages these tails over the chi-square radius, after which the result follows by summing over Hamming shells.

Lemma B.3 (Gaussian radial integration). *There exists a numerical constant $C > 0$ such that, for $W \sim \chi^2_k$ and $\ell, \mu, r > 0$,*

$$\mathbb{E}\exp\big[-\mu\min(W/r^2, W^2/(\ell r^4))\big] \leq \big[Cr\big((\ell/k)^{1/4}\mu^{-1/4} + \mu^{-1/2}\big)\big]^k.$$

B.1 Proof of Proposition 2.1

Let $\mathcal{E}_{\mathrm{des}}$ be the intersection of the event in Lemma A.3 with $\tau = c_0/4$ and the full-design upper-bound event in Lemma A.2(i) with $t = n$. Then

$$\mathbb{P}(\mathcal{E}_{\mathrm{des}}^c) \leq Ce^{-cn}.$$

All deterministic design comparisons below are made on $\mathcal{E}_{\mathrm{des}}$. For $a, b \in \{+,-\}$, define $\mathcal{I}_{ab} := \{i : z_i^* = a,\ z_i = b\}$.

B.1.1 Signal lower bound

For a fixed partition $\mathbf{z}$, the RSS decomposes over the two groups, giving

$$d_\mathbf{z}^2 = \sum_{b\in\{+,-\}} \min_{\mathbf{a}\in\mathbb{R}^p} \big(\|\mathbf{U}_{\mathcal{I}_{+b}}(\boldsymbol{\alpha}_+^* - \mathbf{a})\|^2 + \|\mathbf{U}_{\mathcal{I}_{-b}}(\boldsymbol{\alpha}_-^* - \mathbf{a})\|^2\big).$$

For each $b \in \{+,-\}$, choose $L_b, S_b \in \{\mathcal{I}_{+b}, \mathcal{I}_{-b}\}$ such that $\{L_b, S_b\} = \{\mathcal{I}_{+b}, \mathcal{I}_{-b}\}$ and $|L_b| \geq |S_b|$. Since $\mathbf{z} \in \mathcal{T}_{c_0/2}$,

$$|L_b| \geq c_0 n/4.$$

Set $\mathbf{H}_b := \mathbf{U}_{L_b}^\top\mathbf{U}_{L_b}$ and $\mathbf{G}_b := \mathbf{U}_{S_b}^\top\mathbf{U}_{S_b}$. On $\mathcal{E}_{\mathrm{des}}$,

$$\mathbf{H}_b \succeq cn\mathbf{I}_p, \qquad \mathbf{G}_b \preceq Cn\mathbf{I}_p,$$

and hence $\mathbf{H}_b \succeq c\mathbf{G}_b$.

Since $\boldsymbol{\alpha}_+^* - \boldsymbol{\alpha}_-^* = \Lambda\mathbf{v}$, each groupwise term is of the form in Fact B.2 with $\mathbf{H} = \mathbf{H}_b$, $\mathbf{G} = \mathbf{G}_b$, and $\mathbf{d} = \pm\Lambda\mathbf{v}$. Hence

$$\min_{\mathbf{a}\in\mathbb{R}^p} \big(\|\mathbf{U}_{L_b}\mathbf{a}\|^2 + \|\mathbf{U}_{S_b}(\mathbf{a} \pm \Lambda\mathbf{v})\|^2\big) \geq c\Lambda^2\mathbf{v}^\top\mathbf{G}_b\mathbf{v}.$$

Summing over $b \in \{+,-\}$ gives

$$d_{\mathbf{z}}^2 \geq c\Lambda^2 \sum_{b\in\{+,-\}} \|\mathbf{U}_{S_b}\mathbf{v}\|^2. \tag{8}$$

Let $\ell_{\mathbf{z}} := |S_+| + |S_-|$. Depending on the choices of S_+ and S_-,

$$\ell_{\mathbf{z}} \in \{|I_+^*|, |I_-^*|, m, n-m\}.$$

If $m < c_0 n$, then

$$|\mathcal{I}_{++}| - |\mathcal{I}_{-+}| = |I_+^*| - m > 0, \qquad |\mathcal{I}_{--}| - |\mathcal{I}_{+-}| = |I_-^*| - m > 0.$$

Thus $S_+ = \mathcal{I}_{-+}$ and $S_- = \mathcal{I}_{+-}$, so $\ell_{\mathbf{z}} = m$. If $m \geq c_0 n$, then each possible value of $\ell_{\mathbf{z}}$ is at least $c_0 n$, since $|I_b^*| \geq c_0 n$ for $b \in \{+,-\}$ and $n - m \geq n/2 \geq c_0 n$.

For fixed $\mathbf{z}$, the sets S_+ and S_- are deterministic and disjoint, and $\mathbf{u}_i^\top \mathbf{v} \overset{\text{iid}}{\sim} N(0,1)$. Therefore,

$$\sum_{b\in\{+,-\}} \|\mathbf{U}_{S_b}\mathbf{v}\|^2 \sim \chi^2_{\ell_{\mathbf{z}}}, \qquad \text{where } \ell_{\mathbf{z}} = \begin{cases} m, & m < c_0 n, \\ \geq c_0 n, & m \geq c_0 n. \end{cases} \tag{9}$$

Since $\chi_k^2 > 0$ almost surely for every $k \geq 1$ and $\mathcal{T}_{c_0/2}$ is finite, (8) and (9) imply that, on $\mathcal{E}_{\text{des}}$, every incorrect partition has strictly larger RSS than the true partition almost surely when $\sigma = 0$. Together with $\mathbb{P}(\mathcal{E}_{\text{des}}^c) \leq Ce^{-cn}$, this proves the proposition for $\sigma = 0$. Henceforth assume $\sigma > 0$.

B.1.2 Noise fluctuations for a fixed partition

Fix $\mathbf{z} \neq \mathbf{z}^*$ and condition on $\mathbf{U} \in \mathcal{E}_{\text{des}}$. We first control $\mathbf{P}_{\mathbf{z}} - \mathbf{P}_*$, which determines the quadratic noise fluctuation.

Let $\mathcal{S}_* := \text{col}(\mathbf{D}_*)$ and $\mathcal{S}_{\mathbf{z}} := \text{col}(\mathbf{D}_{\mathbf{z}})$. On $\mathcal{E}_{\text{des}}$, both subspaces have dimension $2p$ and $\sigma_{\min}^2(\mathbf{D}_*) \geq cn$. Since $\mathbf{D}_*(\mathbf{D}_*^\top \mathbf{D}_*)^{-1/2}$ has orthonormal columns spanning $\mathcal{S}_*$,

$$\begin{aligned} \|\sin\Theta(\mathcal{S}_*, \mathcal{S}_{\mathbf{z}})\|_{\text{F}} &= \|(\mathbf{I} - \mathbf{P}_{\mathbf{z}})\mathbf{D}_*(\mathbf{D}_*^\top \mathbf{D}_*)^{-1/2}\|_{\text{F}} \\ &\leq \frac{\|(\mathbf{I} - \mathbf{P}_{\mathbf{z}})\mathbf{D}_*\|_{\text{F}}}{\sigma_{\min}(\mathbf{D}_*)} \\ &\leq \frac{\|\mathbf{D}_* - \mathbf{D}_{\mathbf{z}}\|_{\text{F}}}{\sigma_{\min}(\mathbf{D}_*)}, \end{aligned}$$

where the last inequality uses $(\mathbf{I} - \mathbf{P}_{\mathbf{z}})\mathbf{D}_{\mathbf{z}} = 0$. Also,

$$\|\mathbf{D}_* - \mathbf{D}_{\mathbf{z}}\|_{\text{F}}^2 = 2 \sum_{i: z_i \neq z_i^*} \|\mathbf{u}_i\|^2 \leq 2m\|\mathbf{U}\|_{\text{op}}^2 \leq Cnm.$$

Therefore,

$$\|\mathbf{P}_{\mathbf{z}} - \mathbf{P}_*\|_{\text{op}} \leq 1, \qquad \|\mathbf{P}_{\mathbf{z}} - \mathbf{P}_*\|_{\text{F}}^2 = 2\|\sin\Theta(\mathcal{S}_*, \mathcal{S}_{\mathbf{z}})\|_{\text{F}}^2 \leq Cm, \tag{10}$$

where the operator-norm bound holds for any two orthogonal projectors, and the equality uses that $\mathcal{S}_*$ and $\mathcal{S}_{\mathbf{z}}$ have the same dimension.

Conditional on $\mathbf{U}$, the linear term in (7) is centered Gaussian with variance $4\sigma^2 d_{\mathbf{z}}^2$. Thus

$$\mathbb{P}\big(2\boldsymbol{\epsilon}^\top(\mathbf{I} - \mathbf{P}_{\mathbf{z}})\boldsymbol{\nu} \leq -d_{\mathbf{z}}^2/2 \mid \mathbf{U}\big) \leq \exp(-cd_{\mathbf{z}}^2/\sigma^2).$$

Moreover, $\text{tr}(\mathbf{P}_* - \mathbf{P}_{\mathbf{z}}) = 0$. Conditionally on $\mathbf{U}$, the Hanson–Wright inequality (Vershynin, 2018, Theorem 6.2.1), together with (10), gives

$$\mathbb{P}\big(|\boldsymbol{\epsilon}^\top(\mathbf{P}_* - \mathbf{P}_{\mathbf{z}})\boldsymbol{\epsilon}| \geq d_{\mathbf{z}}^2/2 \mid \mathbf{U}\big) \leq 2\exp\big[-c\min(d_{\mathbf{z}}^2/\sigma^2, d_{\mathbf{z}}^4/(\sigma^4 m))\big].$$

If neither event occurs, then $\text{RSS}(\mathbf{z}) > \text{RSS}(\mathbf{z}^*)$. Therefore,

$$\mathbb{P}\big(\text{RSS}(\mathbf{z}) \leq \text{RSS}(\mathbf{z}^*) \mid \mathbf{U}\big) \leq 3\exp\big[-c\min(d_{\mathbf{z}}^2/\sigma^2, d_{\mathbf{z}}^4/(\sigma^4 m))\big]. \tag{11}$$

B.1.3 HAMMING-SHELL SUMMATION

For $1 \le m \le n/2$, let

$$\mathcal{Z}_m := \{\mathbf{z} \in \mathcal{T}_{c_0/2} : d_H(\mathbf{z}, \mathbf{z}^*) = m\},$$

and write $M := n\bar{h}(\hat{\mathbf{z}}^{\mathrm{prof}}, \mathbf{z}^*)$. Then

$$\{M = m\} \subseteq \bigcup_{\mathbf{z} \in \mathcal{Z}_m} \{\mathrm{RSS}(\mathbf{z}) \le \mathrm{RSS}(\mathbf{z}^*)\}.$$

Fix $\mathbf{z} \in \mathcal{Z}_m$ and set $W_{\mathbf{z}} := \sum_{b \in \{+,-\}} \|\mathbf{U}_{S_b}\mathbf{v}\|^2$. It follows that

$$\begin{aligned}
&\mathbb{P}\big(\mathrm{RSS}(\mathbf{z}) \le \mathrm{RSS}(\mathbf{z}^*), \mathcal{E}_{\mathrm{des}}\big) \\
&\quad = \mathbb{E}\big[\mathbf{I}(\mathcal{E}_{\mathrm{des}})\mathbb{P}(\mathrm{RSS}(\mathbf{z}) \le \mathrm{RSS}(\mathbf{z}^*) \mid \mathbf{U})\big] \\
&\quad \le 3\mathbb{E}\big[\mathbf{I}(\mathcal{E}_{\mathrm{des}}) \exp\big(-c\min(W_{\mathbf{z}}/r^2, W_{\mathbf{z}}^2/(mr^4))\big)\big] \\
&\quad \le 3\mathbb{E}_{W \sim \chi^2_{\ell_{\mathbf{z}}}}\big[\exp\big(-c\min(W/r^2, W^2/(mr^4))\big)\big] \\
&\quad \le \big[Cr\big((m/\ell_{\mathbf{z}})^{1/4} + 1\big)\big]^{\ell_{\mathbf{z}}},
\end{aligned}$$

where the first inequality uses (8) and (11), the second uses (9), and the last follows from Lemma B.3. Since $m/\ell_{\mathbf{z}} \le 1/(2c_0)$, the preceding bound yields

$$\mathbb{P}(\mathrm{RSS}(\mathbf{z}) \le \mathrm{RSS}(\mathbf{z}^*), \mathcal{E}_{\mathrm{des}}) \le (Cr)^{\ell_{\mathbf{z}}}. \tag{12}$$

Set $C_0 := eC$ and choose $r_0 > 0$ such that $2(C_0 r_0)^{c_0} < 1$. For $r > r_0$, both conclusions are immediate from $\bar{h} \le 1$ and $\mathbb{P}(M \ge 1) \le 1$ after enlarging the constants.

It remains to consider $0 < r \le r_0$. Then (12) gives

$$\begin{aligned}
\mathbb{P}(M = m, \mathcal{E}_{\mathrm{des}}) &\le \min\big(1, (C_0 nr/m)^m\big), \quad 1 \le m < c_0 n, \\
\mathbb{P}(M \ge c_0 n, \mathcal{E}_{\mathrm{des}}) &\le 2^n (C_0 r)^{c_0 n} \le e^{-cn},
\end{aligned}$$

where the first bound uses $|\mathcal{Z}_m| \le \binom{n}{m} \le (en/m)^m$ and $\ell_{\mathbf{z}} = m$ for $\mathbf{z} \in \mathcal{Z}_m$, while the second sums over all candidate partitions and uses $\ell_{\mathbf{z}} \ge c_0 n$ whenever $m \ge c_0 n$.

We first bound the expected mis-clustering error. If $C_0 nr \le 1/2$, then

$$\begin{aligned}
\mathbb{E}[(M/n)\mathbf{I}(\mathcal{E}_{\mathrm{des}})] &\le \frac{1}{n} \sum_{1 \le m < c_0 n} m(C_0 nr/m)^m + e^{-cn} \\
&\le \frac{1}{n} \sum_{m \ge 1} m(C_0 nr)^m + e^{-cn} \le Cr + e^{-cn}.
\end{aligned}$$

If $C_0 nr > 1/2$, set $a_0 := eC_0 nr$. Then

$$\begin{aligned}
\mathbb{E}[(M/n)\mathbf{I}(\mathcal{E}_{\mathrm{des}})] &\le \frac{a_0}{n} + \frac{1}{n} \sum_{a_0 \le m < c_0 n} m(C_0 nr/m)^m + \mathbb{P}(M \ge c_0 n, \mathcal{E}_{\mathrm{des}}) \\
&\le eC_0 r + \frac{1}{n} \sum_{m \ge 1} me^{-m} + e^{-cn} \\
&\le Cr + e^{-cn},
\end{aligned}$$

where the last inequality uses $1/n < 2C_0 r$.

Adding $\mathbb{P}(\mathcal{E}_{\mathrm{des}}^c)$ proves the mis-clustering bound.

For exact recovery, if $C_0 nr \le 1/2$, then

$$\begin{aligned}
\mathbb{P}(M \ge 1) &\le \sum_{1 \le m < c_0 n} (C_0 nr/m)^m + \mathbb{P}(M \ge c_0 n, \mathcal{E}_{\mathrm{des}}) + \mathbb{P}(\mathcal{E}_{\mathrm{des}}^c) \\
&\le Cnr + Ce^{-cn}.
\end{aligned}$$

If $C_0 nr > 1/2$, the bound follows from $\mathbb{P}(M \ge 1) \le 1 \le 2C_0 nr$. This completes the proof of the exact-recovery bound in Proposition 2.1.

B.2 Proof of Lemma B.3

By the χ^2_k density, the change of variables $w = r^2 x$, and $e^{-r^2 x/2} \le 1$,

$$\mathbb{E}\exp\bigl[-\mu \min(W/r^2, W^2/(\ell r^4))\bigr] \le \frac{r^k}{2^{k/2}\Gamma(k/2)} \int_0^\infty x^{k/2-1} e^{-\mu \min(x, x^2/\ell)}\, dx.$$

Splitting at $x = \ell$ and applying the substitutions $s = \mu x^2/\ell$ and $s = \mu x$ on the two ranges, respectively, gives

$$\int_0^\ell x^{k/2-1} e^{-\mu x^2/\ell}\, dx = \frac{1}{2}\left(\frac{\ell}{\mu}\right)^{k/4} \int_0^{\mu\ell} s^{k/4-1} e^{-s}\, ds \le \frac{1}{2}\ell^{k/4}\mu^{-k/4}\Gamma(k/4),$$
$$\int_\ell^\infty x^{k/2-1} e^{-\mu x}\, dx = \mu^{-k/2} \int_{\mu\ell}^\infty s^{k/2-1} e^{-s}\, ds \le \mu^{-k/2}\Gamma(k/2).$$

Therefore,

$$\mathbb{E}\bigl[\exp(-\mu \min(W/r^2, W^2/(\ell r^4)))\bigr] \le r^k \left(\frac{\ell^{k/4}\mu^{-k/4}\Gamma(k/4)}{2^{k/2+1}\Gamma(k/2)} + (2\mu)^{-k/2} \right).$$

By Stirling's formula,

$$\frac{\ell^{k/4}\Gamma(k/4)}{2^{k/2}\Gamma(k/2)} \le [C(\ell/k)^{1/4}]^k,$$

and hence

$$\mathbb{E}\bigl[\exp(-\mu \min(W/r^2, W^2/(\ell r^4)))\bigr] \le \bigl[Cr\bigl((\ell/k)^{1/4}\mu^{-1/4} + \mu^{-1/2}\bigr)\bigr]^k.$$

C Information-theoretic lower bounds

This section proves Proposition 2.4 and completes Theorem 1.1. We use the oracle experiment (1) to construct a product subproblem whose MAP error count is binomial. The same construction yields lower bounds for expected Hamming error, exact recovery, and finite-sample Hamming tails.

C.1 Proof of Lemma 2.3

If $r = 0$, then $y = z\Lambda G/2$ and $G \neq 0$ almost surely, so the Bayes error is zero. Now let $r > 0$. Conditional on $G = g$, the two response distributions corresponding to $z = \pm 1$ are $N(\pm\Lambda g/2, \sigma^2)$. Under the uniform prior, the conditional Bayes error is $\Phi(-|g|/(2r))$. Thus, for an independent $H \sim N(0,1)$,

$$q(r) = \mathbb{E}[\Phi(-|G|/(2r))] = \mathbb{P}(H \le -|G|/(2r)).$$

The event on the right is the downward wedge shown in Figure 5, with total opening angle $2\arctan(2r)$. Rotational invariance therefore gives

$$q(r) = \frac{1}{\pi}\arctan(2r) \asymp \min(1, r).$$

C.2 Proof of Proposition 2.4

Proof. Restrict to

$$\boldsymbol{\Sigma} = \mathbf{I}_p, \qquad \boldsymbol{\beta}^*_\pm = \pm\frac{\Lambda}{2}\mathbf{e}_1.$$

Then $y_i = z_i\Lambda x_{i1}/2 + \sigma\xi_i$, with $x_{i1}, \xi_i \overset{\text{iid}}{\sim} N(0,1)$. Reveal the regression parameters, so that each pair (x_{i1}, y_i) has the oracle distribution in (1).

Choose $F \subset [n]$ with $|F| = \lfloor \zeta n \rfloor$, where $0 < \zeta < \min(1/3, 1 - 2c_0)$, and fix $\mathbf{z}^0 \in \{+1, -1\}^{F^c}$ whose two group sizes differ by at most one. For each $\mathbf{a} \in \{+1, -1\}^F$, let $\mathbf{z}^{\mathbf{a}}$ agree with $\mathbf{z}^0$ on F^c and with $\mathbf{a}$ on F. Since $\zeta < 1 - 2c_0$, for all sufficiently large n the fixed block contains at least

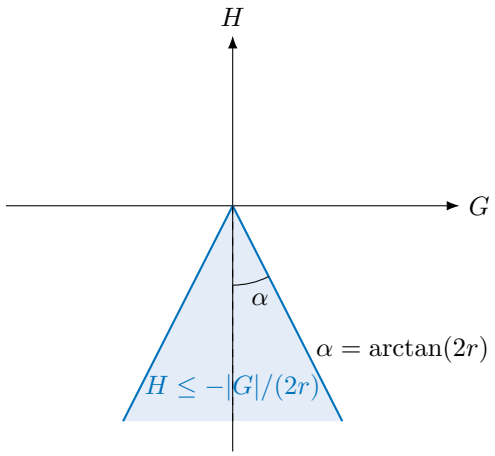


Figure 5: The oracle Bayes error is the probability of the shaded wedge under the standard bivariate Gaussian distribution. Rotational invariance gives $q(r) = \alpha/\pi$.

$c_0 n$ labels of each sign. Hence every $\mathbf{z}^{\mathbf{a}}$ satisfies the balance condition, and these choices form a subfamily of $\Theta_{c_0}(n, p; \Lambda, \sigma)$.

For the lower bound, reveal the labels on F^c and place the uniform prior on the free labels on F. For any estimator $\mathbf{s}$, choose $\widetilde{\mathbf{s}} \in \{\mathbf{s}, -\mathbf{s}\}$ such that $d_H(\widetilde{\mathbf{s}}_{F^c}, \mathbf{z}^0) \le |F^c|/2$, and set $\widehat{\mathbf{a}} := \widetilde{\mathbf{s}}_F$. Then, for every $\mathbf{a} \in \{+1, -1\}^F$,

$$d_H(\widetilde{\mathbf{s}}, \mathbf{z}^{\mathbf{a}}) \ge d_H(\widehat{\mathbf{a}}, \mathbf{a}),$$

while

$$\begin{aligned} d_H(-\widetilde{\mathbf{s}}, \mathbf{z}^{\mathbf{a}}) &\ge |F^c| - d_H(\widetilde{\mathbf{s}}_{F^c}, \mathbf{z}^0) \\ &\ge \frac{|F^c|}{2} \ge |F| \ge d_H(\widehat{\mathbf{a}}, \mathbf{a}), \end{aligned}$$

where $|F| \le n/3$. Therefore

$$n\bar{h}(\mathbf{s}, \mathbf{z}^{\mathbf{a}}) \ge d_H(\widehat{\mathbf{a}}, \mathbf{a}).$$

Thus, any lower bound for recovering the free-label vector $\mathbf{a}$ in the revealed oracle experiment yields a corresponding lower bound for the original mis-clustering problem.

Conditional on the sample $(\mathbf{x}_i, y_i)_{i\in[n]}$ and, if present, the auxiliary randomization of the estimator, $\widehat{\mathbf{a}}$ is fixed while the posterior free labels remain independent. Hence the error indicators $\mathbf{I}(\widehat{a}_i \ne a_i)$, $i \in F$, are independent Bernoulli variables, whose error probabilities are minimized coordinatewise by the MAP rule. Therefore, for every integer $0 \le k \le |F|$,

$$\mathbb{P}\big(d_H(\widehat{\mathbf{a}}, \mathbf{a}) \ge k \mid (x_{i1}, y_i)_{i\in F}\big) \ge \mathbb{P}\big(d_H(\widehat{\mathbf{a}}^{\mathrm{MAP}}, \mathbf{a}) \ge k \mid (x_{i1}, y_i)_{i\in F}\big).$$

Under the uniform product prior on $\mathbf{a} \in \{+1, -1\}^F$, the coordinate experiments are independent across $i \in F$. Hence the MAP error indicators $\mathbf{I}(\widehat{a}_i^{\mathrm{MAP}} \ne a_i)$, $i \in F$, are independent, each with probability $q(r)$ by Lemma 2.3. Therefore

$$d_H(\widehat{\mathbf{a}}^{\mathrm{MAP}}, \mathbf{a}) \overset{\mathrm{d}}{=} B_{|F|,q(r)},$$

where $B_{|F|,q(r)} \sim \mathrm{Bin}(|F|, q(r))$.

Under the same prior, for any membership estimator $\mathbf{s}$,

$$\begin{aligned} \mathbb{E}\bar{h}(\mathbf{s}, \mathbf{z}^{\mathbf{a}}) &\ge \frac{1}{n}\mathbb{E}d_H(\widehat{\mathbf{a}}, \mathbf{a}) \ge \frac{1}{n}\mathbb{E}B_{|F|,q(r)} = \frac{|F|}{n}q(r), \\ \mathbb{P}\big(\bar{h}(\mathbf{s}, \mathbf{z}^{\mathbf{a}}) > 0\big) &\ge \mathbb{P}\big(d_H(\widehat{\mathbf{a}}, \mathbf{a}) > 0\big) \ge \mathbb{P}(B_{|F|,q(r)} > 0) \\ &= 1 - (1 - q(r))^{|F|}. \end{aligned}$$

Taking the infimum over $\mathbf{s}$ and applying the Bayes lower bound on the above subfamily gives

$$R_H^*(n,p,\Lambda,\sigma) \geq \frac{|F|}{n} q(r), \qquad R_{\mathrm{ex}}^*(n,p,\Lambda,\sigma) \geq 1-(1-q(r))^{|F|}.$$

Since $|F| = \lfloor \zeta n \rfloor$, this proves Proposition 2.4. □

C.3 Finite-sample tail lower bound

The preceding argument also gives the following tail lower bound, which will be used in Appendix D to establish the matching lower bound at the scale $\Psi_{n,r}(t)$ when $nr \gtrsim 1$.

Proposition C.1. *Let $m := \lfloor \zeta n \rfloor$ as in the proof of Proposition 2.4. Then, for all sufficiently large n and every integer $0 \leq k \leq m$,*

$$\inf_{\widehat{\mathbf{z}}} \sup_{\vartheta \in \Theta_{c_0}(n,p;\Lambda,\sigma)} \mathbb{P}_\vartheta\big(n\bar{h}(\widehat{\mathbf{z}},\mathbf{z}) \geq k\big) \geq \mathbb{P}\big(B_{m,q(r)} \geq k\big),$$

where $B_{m,q(r)} \sim \mathrm{Bin}(m, q(r))$.

C.4 Proof of Theorem 1.1

For mis-clustering, Proposition 2.4 and Lemma 2.3 give $R_H^*(n,p,\Lambda,\sigma) \gtrsim \min(1,r)$, while Proposition 2.1 gives the corresponding upper bound. This proves the finite-sample statement for all sufficiently large n.

For exact recovery, the upper bound in Proposition 2.1 gives sufficiency of $nr \to 0$. Conversely, if $R_{\mathrm{ex}}^* \to 0$, the oracle lower bound implies

$$(1-q(r))^{\lfloor \zeta n \rfloor} \to 1.$$

Let $m = \lfloor \zeta n \rfloor$. Since

$$1 \geq e^{-mq(r)} \geq (1-q(r))^m \to 1,$$

we have $mq(r) \to 0$. Since $m \asymp n$, this implies $q(r) \to 0$ and hence $r \to 0$. Using $q(r)/r \to 2/\pi$ then gives $nr \to 0$, proving the exact-recovery statement.

D Proofs for local membership contraction

This appendix proves Theorem 4.1 and Corollary 4.3. We first introduce the notation and simple facts shared by the proofs.

Gaussian coordinates and bounds. Using the whitening and normalized-noise notation of Appendix A, set

$$\widetilde{\mathbf{u}}_i := (\mathbf{u}_i, \xi_i), \qquad \bar{\mathbf{v}} := (\mathbf{v}, 0).$$

Write

$$G_i := \widetilde{\mathbf{u}}_i^\top \bar{\mathbf{v}} = \mathbf{u}_i^\top \mathbf{v}, \qquad \widetilde{\mathbf{u}}_{i,\perp} := \widetilde{\mathbf{u}}_i - G_i \bar{\mathbf{v}}.$$

Let $\mathcal{E}_{\mathrm{G}}(t)$ denote the event in Corollary A.4. With ψ_n from Appendix A, define

$$\varepsilon_n(t) := \psi_n(p,t) \log \frac{e}{\psi_n(p,t)}.$$

Under $p + t + \log(en) \leq \kappa n$, both $\psi_n(p,t)$ and $\varepsilon_n(t)$ can be made arbitrarily small by taking κ sufficiently small.

We will repeatedly use the following elementary Gaussian identities and bounds.

Fact D.1. *For $G \sim N(0,1)$, $u, a \geq 0$, and every integer $q \geq 1$,*

$$\mathbb{P}(|G| \leq u) \leq u, \qquad \mathbb{E}e^{-aG^2} = (1+2a)^{-1/2},$$
$$\mathbb{E}[G^2 \mathbf{I}(|G| \geq u)] \leq 2\sqrt{2}\, e^{-u^2/4}, \qquad \mathbb{E}|G|^{2q} \leq 2^q q!.$$

Proof. The first two bounds follow directly from the standard Gaussian density. For the truncated second moment,

$$\mathbb{E}[G^2\mathbf{I}(|G| \geq u)] \leq e^{-u^2/4}\mathbb{E}[G^2 e^{G^2/4}] = 2\sqrt{2}\, e^{-u^2/4}.$$

For the even moments,

$$\mathbb{E}|G|^{2q} = (2q-1)!! \leq 2^q q!.$$

□

Assignments and least-squares refits. We next introduce the assignment sets and least-squares refits induced by a local coefficient pair.

Fix a local coefficient pair γ, oriented by the minimizing permutation in $a(\gamma)$ as in Section 4, and set

$$s_\gamma := r + a(\gamma).$$

The nearest-residual rule induced by γ gives the assignments $\widehat{z}_i(\gamma)$. For $b \in \{+,-\}$, let

$$\widehat{I}_b(\gamma) := \{i : \widehat{z}_i(\gamma) = b\}, \qquad E_{\gamma,b} := \{i \in I_b^* : \widehat{z}_i(\gamma) \neq b\}.$$

Thus $\widehat{I}_b(\gamma)$ is the set assigned to component b, whereas $E_{\gamma,b}$ consists of observations from true group b that are misassigned by γ. Write

$$E_\gamma := E_{\gamma,+} \cup E_{\gamma,-}, \qquad \mathbf{B}_{\gamma,b} := \frac{1}{n}\sum_{i \in E_{\gamma,b}} \mathbf{u}_i\mathbf{u}_i^\top,$$

for the full misassignment set and the normalized selected second-moment matrix over the misassigned observations from true group b, respectively. The assignment $\widehat{I}_b(\gamma)$ also determines the normalized Gram matrix

$$\mathbf{S}_{\gamma,b} := \frac{1}{n}\sum_{i \in \widehat{I}_b(\gamma)} \mathbf{u}_i\mathbf{u}_i^\top.$$

For comparison, the corresponding normalized true-group Gram matrix is

$$\mathbf{S}_b^* := \frac{1}{n}\sum_{i \in I_b^*} \mathbf{u}_i\mathbf{u}_i^\top.$$

Finally, refitting least squares on the groups determined by γ gives

$$\mathcal{M}_b(\gamma) := \Big(\sum_{i \in \widehat{I}_b(\gamma)} \mathbf{x}_i\mathbf{x}_i^\top\Big)^{-1} \sum_{i \in \widehat{I}_b(\gamma)} \mathbf{x}_i y_i, \qquad b \in \{+,-\},$$

and

$$\mathcal{M}(\gamma) := \big(\mathcal{M}_+(\gamma), \mathcal{M}_-(\gamma)\big).$$

Fact D.2. *For every coefficient pair γ and $b \in \{+,-\}$,*

$$\widehat{I}_b(\gamma) = (I_b^* \setminus E_{\gamma,b}) \cup E_{\gamma,-b}, \qquad \mathbf{S}_{\gamma,b} = \mathbf{S}_b^* - \mathbf{B}_{\gamma,b} + \mathbf{B}_{\gamma,-b}.$$

For every $\boldsymbol{\alpha} \in \mathbb{R}^p$,

$$\mathbf{S}_{\gamma,b}\frac{\mathbf{A}\mathcal{M}_b(\gamma) - \boldsymbol{\alpha}}{\Lambda} = \frac{1}{n\Lambda}\sum_{i \in \widehat{I}_b(\gamma)} \mathbf{u}_i(y_i - \mathbf{u}_i^\top\boldsymbol{\alpha}). \tag{13}$$

Proof. The decomposition of $\widehat{I}_b(\gamma)$ follows directly from the definitions. Substituting it into the definition of $\mathbf{S}_{\gamma,b}$ gives the second decomposition. Using $\mathbf{X}_{\widehat{I}_b(\gamma)} = \mathbf{U}_{\widehat{I}_b(\gamma)}\mathbf{A}$, the least-squares normal equation gives

$$\mathbf{S}_{\gamma,b}\mathbf{A}\mathcal{M}_b(\gamma) = \frac{1}{n}\sum_{i \in \widehat{I}_b(\gamma)} \mathbf{u}_i y_i.$$

Since $\mathbf{S}_{\gamma,b}\boldsymbol{\alpha} = \frac{1}{n}\sum_{i \in \widehat{I}_b(\gamma)} \mathbf{u}_i\mathbf{u}_i^\top\boldsymbol{\alpha}$, (13) follows. □

Wedge representation. The nearest-residual comparisons defining $E_{\gamma,b}$ can be written as sign-disagreement events. The following deterministic normalization represents such events as wedges around the margin direction $\bar{\mathbf{v}}$.

Fact D.3. *Let* $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^{p+1}$ *satisfy* $\bar{\mathbf{v}}^\top \mathbf{a}_j > 0$*, and define*

$$E := \{(\mathbf{a}_1^\top \widetilde{\mathbf{u}})(\mathbf{a}_2^\top \widetilde{\mathbf{u}}) \le 0\}.$$

For $j \in \{1, 2\}$*, set*

$$c_j := \bar{\mathbf{v}}^\top \mathbf{a}_j, \qquad \mathbf{h}_j := \frac{\mathbf{a}_j}{c_j} - \bar{\mathbf{v}}.$$

Then $\mathbf{h}_j \perp \bar{\mathbf{v}}$ *and*

$$E = \{(G + \mathbf{h}_1^\top \widetilde{\mathbf{u}}_\perp)(G + \mathbf{h}_2^\top \widetilde{\mathbf{u}}_\perp) \le 0\}. \tag{14}$$

Moreover,

$$G^2 \mathbf{I}(E) \le \sum_{j=1}^{2} (\mathbf{h}_j^\top \widetilde{\mathbf{u}})^2. \tag{15}$$

Proof. By the definitions of c_j and $\mathbf{h}_j$, $\bar{\mathbf{v}}^\top \mathbf{h}_j = 0$ and $\mathbf{a}_j = c_j(\bar{\mathbf{v}} + \mathbf{h}_j)$. Hence $\mathbf{a}_j^\top \widetilde{\mathbf{u}} = c_j(G + \mathbf{h}_j^\top \widetilde{\mathbf{u}}_\perp)$, so (14) follows since $c_1 c_2 > 0$.

On E, $-G$ lies between $\mathbf{h}_1^\top \widetilde{\mathbf{u}}_\perp$ and $\mathbf{h}_2^\top \widetilde{\mathbf{u}}_\perp$. Therefore

$$G^2 \le \sum_{j=1}^{2} (\mathbf{h}_j^\top \widetilde{\mathbf{u}}_\perp)^2 = \sum_{j=1}^{2} (\mathbf{h}_j^\top \widetilde{\mathbf{u}})^2,$$

where the equality uses $\mathbf{h}_j \perp \bar{\mathbf{v}}$. □

Weighted and ordinary Hamming losses. Recall $H(\gamma)$ and $H^{\mathrm{w}}(\gamma)$ from Section 4. By the definitions of $E_{\gamma,b}$,

$$H^{\mathrm{w}}(\gamma) = \frac{1}{n} \sum_{i \in E_\gamma} G_i^2 = \frac{1}{n} \sum_{b \in \{+,-\}} \sum_{i \in E_{\gamma,b}} G_i^2. \tag{16}$$

The weighted loss controls the ordinary Hamming loss through the following margin decomposition.

Fact D.4. *For every coefficient pair* γ *and* $h > 0$*,*

$$H(\gamma) \le \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(|G_i| \le h) + \frac{H^{\mathrm{w}}(\gamma)}{h^2}.$$

Proof.

$$\mathbf{I}(\widehat{z}_i(\gamma) \ne z_i^*) \le \mathbf{I}(|G_i| \le h) + \frac{G_i^2}{h^2} \mathbf{I}(\widehat{z}_i(\gamma) \ne z_i^*).$$

Summing over i gives the result. □

Oracle estimators and prediction errors. For $b \in \{+, -\}$, define the oracle least-squares estimator by

$$\boldsymbol{\alpha}_b^\circ := (\mathbf{S}_b^*)^{-1} \frac{1}{n} \sum_{i \in I_b^*} \mathbf{u}_i y_i.$$

Its normalized estimation and prediction errors are

$$\mathbf{d}_b^\circ := \frac{\boldsymbol{\alpha}_b^\circ - \boldsymbol{\alpha}_b^*}{\Lambda}, \qquad Z_{ib}^\circ := \mathbf{u}_i^\top \mathbf{d}_b^\circ.$$

Define also the normalized displacement of the refitted coefficient from the oracle estimator and the normalized oracle residual by

$$\mathbf{d}_b(\gamma) := \frac{\mathbf{A}\mathcal{M}_b(\gamma) - \boldsymbol{\alpha}_b^\circ}{\Lambda}, \qquad r_{ib}^\circ := \frac{y_i - \mathbf{u}_i^\top \boldsymbol{\alpha}_b^\circ}{\Lambda}.$$

Here $\mathbf{d}_b^\circ$ is the oracle estimation error, while $\mathbf{d}_b(\gamma)$ is the additional displacement induced by the current assignments. We will use the following identities.

Fact D.5. *For every $b \in \{+,-\}$,*

$$\mathbf{d}_b^\circ = r(\mathbf{S}_b^*)^{-1}\frac{1}{n}\sum_{i\in I_b^*}\mathbf{u}_i\xi_i. \tag{17}$$

For every coefficient pair γ,

$$\mathbf{S}_{\gamma,b}\mathbf{d}_b(\gamma) = -\frac{1}{n}\sum_{i\in E_{\gamma,b}}\mathbf{u}_i r_{ib}^\circ + \frac{1}{n}\sum_{i\in E_{\gamma,-b}}\mathbf{u}_i r_{ib}^\circ. \tag{18}$$

Moreover,

$$r_{ib}^\circ = \begin{cases} r\xi_i - Z_{ib}^\circ, & i \in I_b^*, \\ \mp G_i + r\xi_i - Z_{ib}^\circ, & i \in I_{-b}^*, \end{cases} \tag{19}$$

where the upper sign corresponds to $b = +$ and the lower sign to $b = -$.

Proof. Equation (17) follows directly from the definition of the oracle estimator.

For the displacement, the oracle normal equation gives $\sum_{i\in I_b^*}\mathbf{u}_i r_{ib}^\circ = 0$. Applying (13) in Fact D.2 with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_b^\circ$, and then using

$$\widehat{I}_b(\gamma) = (I_b^* \setminus E_{\gamma,b}) \cup E_{\gamma,-b},$$

yields

$$\mathbf{S}_{\gamma,b}\mathbf{d}_b(\gamma) = -\frac{1}{n}\sum_{i\in E_{\gamma,b}}\mathbf{u}_i r_{ib}^\circ + \frac{1}{n}\sum_{i\in E_{\gamma,-b}}\mathbf{u}_i r_{ib}^\circ,$$

which proves (18).

Finally, the two cases in (19) follow directly from the model. For $i \in I_b^*$, $r_{ib}^\circ = r\xi_i - Z_{ib}^\circ$, whereas for $i \in I_{-b}^*$, $r_{ib}^\circ = \mp G_i + r\xi_i - Z_{ib}^\circ$, using $\boldsymbol{\alpha}_+^* - \boldsymbol{\alpha}_-^* = \Lambda\mathbf{v}$. □

Oracle-instability remainder. The oracle decomposition separates the refit error into the oracle error $\mathbf{d}_b^\circ$ and the displacement $\mathbf{d}_b(\gamma)$ induced by the current misassignments. The latter will be controlled by $H^{\mathrm{w}}(\gamma)$ through (18). The terms not absorbed into this contraction arise from small realized margins and large oracle prediction errors.

To control the oracle residuals and oracle-instability terms, define

$$w_i := G_i^2 + r^2\xi_i^2, \qquad i \in [n].$$

Fix a sufficiently small numerical constant $\eta > 0$ and define

$$\mathcal{E}_i^{\mathrm{mar}} := \{|G_i| \le 8r|\xi_i|\}, \qquad \mathcal{E}_{ib}^{\mathrm{or}} := \{|Z_{ib}^\circ| > \eta\sqrt{w_i}\}.$$

The first event corresponds to a small realized margin relative to the noise, while the second corresponds to a large oracle prediction error. Define

$$Q^\circ := \frac{1}{n}\sum_{i=1}^n\Big(w_i\mathbf{I}(\mathcal{E}_i^{\mathrm{mar}}) + \sum_{b\in\{+,-\}}\big(w_i + (Z_{ib}^\circ)^2\big)\mathbf{I}(\mathcal{E}_{ib}^{\mathrm{or}})\Big).$$

Thus Q° is sample-dependent but does not depend on the current coefficient pair γ.

Proof idea. Appendix D.1 combines Facts D.2 and D.3 with the Gaussian-wedge bounds in Appendix D.5. Local misassignments form narrow wedges, so their selected second moments are small; this yields well-conditioned assigned-group Gram matrices and stability of one least-squares update.

Appendix D.2 then uses Fact D.5 to express the displacement of the refitted estimator from the oracle estimator through the current misassignments. Outside the exceptional events collected in Q°, new weighted misassignments are controlled by this displacement, giving a one-step contraction of H^{w} up to Q°.

Appendix D.3 bounds Q° by treating the small-margin, cross-group oracle, and in-sample oracle contributions separately; the required Gaussian and leave-one-out arguments are proved in Appendix D.5. Finally, Appendix D.4 iterates the weighted-Hamming contraction and applies Fact D.4 to obtain the Hamming and exact-recovery bounds.

D.1 LOCAL GEOMETRY AND STABILITY OF THE ITERATES

We first establish uniform selected second-moment bounds for local misassignments and well-conditioning of the assigned-group Gram matrices. These bounds then yield stability of one hard least-squares update.

Lemma D.6 (Uniform geometry of local misassignments)**.** *There exist constants $C, c, \rho, r_0, \kappa > 0$, depending only on c_0, such that the following holds. For every $t \geq 1$, if $r \leq r_0$ and $p + t + \log(en) \leq \kappa n$, then with probability at least $1 - Ce^{-t}$, uniformly over $a(\gamma) \leq \rho$ and $b \in \{+, -\}$,*

$$\|\mathbf{B}_{\gamma,b}\|_{\mathrm{op}} \leq C(s_\gamma + \varepsilon_n(t)), \tag{20}$$

$$\mathbf{v}^\top \mathbf{B}_{\gamma,b}\mathbf{v} \leq C s_\gamma^2, \tag{21}$$

$$\|\mathbf{B}_{\gamma,b}\mathbf{v}\|^2 \leq C s_\gamma^2 (s_\gamma + \varepsilon_n(t)), \tag{22}$$

$$\lambda_{\min}(\mathbf{S}_{\gamma,b}) \geq c. \tag{23}$$

Proof. Orient γ by the minimizing permutation in $a(\gamma)$ and denote $\mathbf{f}_b := (\mathbf{A}\boldsymbol{\gamma}_b - \boldsymbol{\alpha}_b^*)/\Lambda$, where $\|\mathbf{f}_b\| \leq a(\gamma)$. For a true-$b$ observation, the normalized residuals under the correct and incorrect components are

$$R_b = r\xi - \mathbf{u}^\top \mathbf{f}_b, \qquad R_{-b} = bG + r\xi - \mathbf{u}^\top \mathbf{f}_{-b}.$$

Hence misassignment occurs when

$$0 \geq R_{-b}^2 - R_b^2 = \big(G + \mathbf{u}^\top(\mathbf{f}_+ - \mathbf{f}_-)\big)\big(G + 2br\xi - b\mathbf{u}^\top(\mathbf{f}_+ + \mathbf{f}_-)\big).$$

Write $\mathbf{f}_b = f_{b,\parallel}\mathbf{v} + \mathbf{f}_{b,\perp}$ with $\mathbf{f}_{b,\perp} \perp \mathbf{v}$, and set

$$\tau_1 := 1 + f_{+,\parallel} - f_{-,\parallel}, \qquad \tau_{2,b} := 1 - b(f_{+,\parallel} + f_{-,\parallel}).$$

Since $|f_{b,\parallel}| \leq \rho$,

$$\tau_1, \tau_{2,b} \geq 1 - 2\rho.$$

Thus, for sufficiently small ρ, Fact D.3 gives the normalized wedge representation with

$$\mathbf{h}_1 := \frac{(\mathbf{f}_{+,\perp} - \mathbf{f}_{-,\perp}, 0)}{\tau_1}, \qquad \mathbf{h}_{2,b} := \frac{(-b(\mathbf{f}_{+,\perp} + \mathbf{f}_{-,\perp}),\, 2br)}{\tau_{2,b}}.$$

Both perturbation vectors are orthogonal to $\bar{\mathbf{v}}$ and satisfy

$$\|\mathbf{h}_1\| + \|\mathbf{h}_{2,b}\| \leq C(r + a(\gamma)) = C s_\gamma.$$

Taking C_0 in the definition of $\mathcal{W}(s)$ in Appendix D.5.1 sufficiently large, the misassignment events for both $b \in \{+, -\}$ belong to $\mathcal{W}(s_\gamma)$. Choose ρ and r_0 so that $s_\gamma \leq s_\star$ throughout the local neighborhood, and then choose κ so that the condition of Lemma D.12 is satisfied.

For the corresponding wedge E, $\mathbf{B}_{\gamma,b} \preceq \mathbf{B}_E$. Hence Lemma D.12 gives (20) and (21) uniformly over $a(\gamma) \leq \rho$ and $b \in \{+, -\}$. Since $\mathbf{B}_{\gamma,b} \succeq 0$,

$$\|\mathbf{B}_{\gamma,b}\mathbf{v}\|^2 \leq \|\mathbf{B}_{\gamma,b}\|_{\mathrm{op}} \mathbf{v}^\top \mathbf{B}_{\gamma,b}\mathbf{v},$$

which yields (22).

Intersect the event above with $\mathcal{E}_{\mathrm{G}}(t)$. By the Gram identity in Fact D.2 and the true-group Gram bound (3), $\mathbf{S}_b^* \succeq c\mathbf{I}_p$. Since $\mathbf{B}_{\gamma,-b} \succeq 0$,

$$\lambda_{\min}(\mathbf{S}_{\gamma,b}) \geq c - \|\mathbf{B}_{\gamma,b}\|_{\mathrm{op}}.$$

Choose ρ sufficiently small, then r_0, and finally κ, so that

$$C\big(s_\gamma + \varepsilon_n(t)\big) \leq c/2$$

uniformly over $a(\gamma) \leq \rho$. Together with (20), this proves (23). A union bound gives the stated probability. □

The preceding geometry bounds yield the following stability result for one hard least-squares update.

Lemma D.7 (Local-neighborhood stability)**.** *Under the conditions of Lemma D.6, there is an event $\mathcal{G}_{\mathrm{loc}}(t)$ of probability at least $1 - Ce^{-t}$ such that, uniformly over $a(\gamma) \le \rho$,*

$$a(\mathcal{M}(\gamma)) \le C\left(s_\gamma\sqrt{s_\gamma + \varepsilon_n(t)} + r\sqrt{\frac{p+t}{n}}\right) \le \frac{1}{2}a(\gamma) + Cr. \tag{24}$$

The constants can be chosen with $Cr_0 \le \rho/4$, so that $\mathcal{M}$ maps the ρ-neighborhood into itself. In particular, any possibly data-dependent $\gamma^{(0)}$ with $a(\gamma^{(0)}) \le \rho/2$ satisfies

$$a(\gamma^{(k)}) \le \rho, \qquad k \ge 0.$$

Proof. Let $\mathcal{G}_{\mathrm{loc}}(t)$ be the intersection of the event in Lemma D.6 and $\mathcal{E}_{\mathrm{G}}(t)$. Then $\mathbb{P}\big(\mathcal{G}_{\mathrm{loc}}^c(t)\big) \le Ce^{-t}$. For $b \in \{+,-\}$, take $\boldsymbol{\alpha} = \boldsymbol{\alpha}_b^*$ in (13) of Fact D.2. Since $\boldsymbol{\alpha}_+^* - \boldsymbol{\alpha}_-^* = \Lambda\mathbf{v}$, this gives

$$\mathbf{S}_{\gamma,b}\frac{\mathbf{A}\mathcal{M}_b(\gamma) - \boldsymbol{\alpha}_b^*}{\Lambda} = \pm\mathbf{B}_{\gamma,-b}\mathbf{v} + r\frac{1}{n}\sum_{\widehat{z}_i(\gamma)=b}\mathbf{u}_i\xi_i, \tag{25}$$

where the sign is negative for $b = +$ and positive for $b = -$. Using $\widehat{I}_b(\gamma) = (I_b^* \setminus E_{\gamma,b}) \cup E_{\gamma,-b}$,

$$\frac{1}{n}\sum_{\widehat{z}_i(\gamma)=b}\mathbf{u}_i\xi_i = \frac{1}{n}\sum_{i\in I_b^*}\mathbf{u}_i\xi_i - \frac{1}{n}\sum_{i\in E_{\gamma,b}}\mathbf{u}_i\xi_i + \frac{1}{n}\sum_{i\in E_{\gamma,-b}}\mathbf{u}_i\xi_i.$$

Corollary A.4, through (6), gives

$$\Big\|\frac{1}{n}\sum_{i\in I_b^*}\mathbf{u}_i\xi_i\Big\| \le C\sqrt{\frac{p+t}{n}}, \qquad \frac{1}{n}\sum_{i=1}^n \xi_i^2 \le C.$$

By Cauchy–Schwarz and (20),

$$\Big\|\frac{1}{n}\sum_{i\in E_{\gamma,b}}\mathbf{u}_i\xi_i\Big\|^2 \le \|\mathbf{B}_{\gamma,b}\|_{\mathrm{op}}\frac{1}{n}\sum_{i=1}^n\xi_i^2 \le C\big(s_\gamma + \varepsilon_n(t)\big).$$

It follows that, for $b \in \{+,-\}$,

$$\Big\|\frac{1}{n}\sum_{\widehat{z}_i(\gamma)=b}\mathbf{u}_i\xi_i\Big\| \le C\left(\sqrt{\frac{p+t}{n}} + \sqrt{s_\gamma + \varepsilon_n(t)}\right). \tag{26}$$

Lemma D.6, through (22) and (23), gives

$$\|\mathbf{B}_{\gamma,-b}\mathbf{v}\| \le Cs_\gamma\sqrt{s_\gamma + \varepsilon_n(t)}, \qquad \|\mathbf{S}_{\gamma,b}^{-1}\|_{\mathrm{op}} \le C.$$

Combining these bounds with (25) and (26) gives

$$\begin{aligned}\frac{\|\mathbf{A}\mathcal{M}_b(\gamma) - \boldsymbol{\alpha}_b^*\|}{\Lambda} &\le C\left(s_\gamma\sqrt{s_\gamma + \varepsilon_n(t)} + r\sqrt{\frac{p+t}{n}} + r\sqrt{s_\gamma + \varepsilon_n(t)}\right)\\ &\le C\left(s_\gamma\sqrt{s_\gamma + \varepsilon_n(t)} + r\sqrt{\frac{p+t}{n}}\right).\end{aligned}$$

Choose ρ, r_0, and κ sufficiently small, in this order, so that $C\sqrt{s_\gamma + \varepsilon_n(t)} \le \frac{1}{2}$ uniformly over $a(\gamma) \le \rho$. Since $s_\gamma = r + a(\gamma)$ and $(p+t)/n \le \kappa$, the preceding bound gives

$$a(\mathcal{M}(\gamma)) \le \frac{1}{2}a(\gamma) + Cr,$$

which proves (24). If $Cr_0 \le \rho/4$, then $a(\gamma) \le \rho$ implies $a(\mathcal{M}(\gamma)) \le 3\rho/4$. Thus $\mathcal{M}$ maps the ρ-neighborhood into itself, and induction completes the proof. □

D.2 One-step contraction of the weighted Hamming loss

We next show that one hard least-squares update contracts the weighted Hamming loss up to the oracle-instability remainder Q°, using the decomposition in Fact D.5.

Lemma D.8. *On $\mathcal{G}_{\mathrm{loc}}(t)$, uniformly over $a(\gamma) \le \rho$, the update $\gamma' = \mathcal{M}(\gamma)$ satisfies*

$$H^{\mathrm{w}}(\gamma') \le \frac{1}{2} H^{\mathrm{w}}(\gamma) + CQ^\circ. \tag{27}$$

Proof. By (18) in Fact D.5,

$$\mathbf{S}_{\gamma,b}\mathbf{d}_b(\gamma) = \frac{1}{n}\mathbf{U}_{E_\gamma}^\top \mathbf{r}_b,$$

where $\mathbf{U}_{E_\gamma}$ is the whitened design matrix restricted to E_γ, and $\mathbf{r}_b \in \mathbb{R}^{|E_\gamma|}$ is defined by

$$(\mathbf{r}_b)_i = \begin{cases} -r_{ib}^\circ, & i \in E_{\gamma,b}, \\ r_{ib}^\circ, & i \in E_{\gamma,-b}. \end{cases}$$

Lemma D.6 gives $\|\mathbf{S}_{\gamma,b}^{-1}\|_{\mathrm{op}} \le C$, so

$$\|\mathbf{d}_b(\gamma)\|^2 \le C\big(\frac{1}{n}\|\mathbf{U}_{E_\gamma}\|_{\mathrm{op}}^2\big)\big(\frac{1}{n}\|\mathbf{r}_b\|^2\big).$$

By the definition of $\mathbf{B}_{\gamma,b}$ and (20),

$$\frac{1}{n}\|\mathbf{U}_{E_\gamma}\|_{\mathrm{op}}^2 = \|\mathbf{B}_{\gamma,+} + \mathbf{B}_{\gamma,-}\|_{\mathrm{op}} \le C\big(s_\gamma + \varepsilon_n(t)\big).$$

For $i \in E_\gamma$, (19) in Fact D.5 gives

$$(r_{ib}^\circ)^2 \le Cw_i + C\big(w_i + (Z_{ib}^\circ)^2\big)\mathbf{I}(\mathcal{E}_{ib}^{\mathrm{or}}),$$

where $(r_{ib}^\circ)^2 \le Cw_i$ on $(\mathcal{E}_{ib}^{\mathrm{or}})^c$, while $(r_{ib}^\circ)^2 \le C\big(w_i + (Z_{ib}^\circ)^2\big)$ on $\mathcal{E}_{ib}^{\mathrm{or}}$. On $(\mathcal{E}_i^{\mathrm{mar}})^c$, $r^2\xi_i^2 < G_i^2/64$ and $w_i \le 65G_i^2/64$. Therefore, by (16),

$$\frac{1}{n}\sum_{i\in E_\gamma} w_i = \frac{1}{n}\sum_{i\in E_\gamma} w_i \mathbf{I}((\mathcal{E}_i^{\mathrm{mar}})^c) + \frac{1}{n}\sum_{i\in E_\gamma} w_i \mathbf{I}(\mathcal{E}_i^{\mathrm{mar}}) \le \frac{65}{64}H^{\mathrm{w}}(\gamma) + Q^\circ.$$

Hence

$$\frac{1}{n}\|\mathbf{r}_b\|^2 \le \frac{C}{n}\sum_{i\in E_\gamma} w_i + \frac{C}{n}\sum_{i=1}^n \big(w_i + (Z_{ib}^\circ)^2\big)\mathbf{I}(\mathcal{E}_{ib}^{\mathrm{or}}) \le C\big(H^{\mathrm{w}}(\gamma) + Q^\circ\big),$$

where the last inequality follows from the definition of Q°.

Combining the two bounds yields

$$\|\mathbf{d}_b(\gamma)\|^2 \le C\big(s_\gamma + \varepsilon_n(t)\big)\big(H^{\mathrm{w}}(\gamma) + Q^\circ\big). \tag{28}$$

We now control the weighted misassignments after the update. For $b \in \{+,-\}$, define the exceptional event

$$\mathcal{A}_{ib} := \mathcal{E}_i^{\mathrm{mar}} \cup \mathcal{E}_{ib}^{\mathrm{or}} \cup \mathcal{E}_{i,-b}^{\mathrm{or}}.$$

By (16),

$$H^{\mathrm{w}}(\gamma') = \frac{1}{n}\sum_{b\in\{+,-\}}\sum_{i\in E_{\gamma',b}} G_i^2 \mathbf{I}(\mathcal{A}_{ib}) + \frac{1}{n}\sum_{b\in\{+,-\}}\sum_{i\in E_{\gamma',b}} G_i^2 \mathbf{I}(\mathcal{A}_{ib}^c). \tag{29}$$

By the definition of Q° and $G_i^2 \le w_i$, the first term is at most Q°. Thus it remains to control the second term.

Fix $b \in \{+,-\}$ and $i \in E_{\gamma',b}$ with $\mathcal{A}_{ib}^c$. Then $i \in I_b^*$ and

$$r|\xi_i| < \frac{1}{8}|G_i|, \qquad w_i \le 2G_i^2, \qquad |Z_{ib}^\circ|,\ |Z_{i,-b}^\circ| \le 2\eta|G_i|.$$

By (19),

$$|r^\circ_{i,-b}| - |r^\circ_{ib}| \geq |G_i| - 2r|\xi_i| - |Z^\circ_{ib}| - |Z^\circ_{i,-b}|$$
$$\geq \big(\tfrac{3}{4} - 4\eta\big)|G_i| \geq \frac{1}{2}|G_i|,$$

where the last inequality uses $\eta \leq 1/16$.

Since $i \in E_{\gamma',b}$, the nearest-residual rule for $\gamma' = \mathcal{M}(\gamma)$ gives

$$\big|r^\circ_{i,-b} - \mathbf{u}_i^\top \mathbf{d}_{-b}(\gamma)\big| \leq \big|r^\circ_{ib} - \mathbf{u}_i^\top \mathbf{d}_b(\gamma)\big|,$$

where the two sides are the normalized residuals under components $-b$ and b, respectively.

Hence, by the triangle inequality,

$$|r^\circ_{i,-b}| \leq \big|r^\circ_{i,-b} - \mathbf{u}_i^\top \mathbf{d}_{-b}(\gamma)\big| + |\mathbf{u}_i^\top \mathbf{d}_{-b}(\gamma)|$$
$$\leq \big|r^\circ_{ib} - \mathbf{u}_i^\top \mathbf{d}_b(\gamma)\big| + |\mathbf{u}_i^\top \mathbf{d}_{-b}(\gamma)|$$
$$\leq |r^\circ_{ib}| + |\mathbf{u}_i^\top \mathbf{d}_b(\gamma)| + |\mathbf{u}_i^\top \mathbf{d}_{-b}(\gamma)|.$$

Therefore,

$$\frac{1}{2}|G_i| \leq |r^\circ_{i,-b}| - |r^\circ_{ib}| \leq |\mathbf{u}_i^\top \mathbf{d}_b(\gamma)| + |\mathbf{u}_i^\top \mathbf{d}_{-b}(\gamma)|,$$

and we obtain the pointwise bound

$$G_i^2 \leq C\Big[(\mathbf{u}_i^\top \mathbf{d}_b(\gamma))^2 + (\mathbf{u}_i^\top \mathbf{d}_{-b}(\gamma))^2\Big]. \tag{30}$$

By (29) and (30),

$$H^{\mathrm{w}}(\gamma') \leq CQ^\circ + \frac{C}{n} \sum_{b\in\{+,-\}} \sum_{i\in E_{\gamma',b}} \Big[(\mathbf{u}_i^\top \mathbf{d}_b(\gamma))^2 + (\mathbf{u}_i^\top \mathbf{d}_{-b}(\gamma))^2\Big] \mathbf{I}(\mathcal{A}^c_{ib})$$
$$\leq CQ^\circ + \frac{C}{n} \sum_{b\in\{+,-\}} \sum_{i\in E_{\gamma',b}} \Big[(\mathbf{u}_i^\top \mathbf{d}_b(\gamma))^2 + (\mathbf{u}_i^\top \mathbf{d}_{-b}(\gamma))^2\Big]$$
$$= CQ^\circ + C \sum_{b\in\{+,-\}} \mathbf{d}_b(\gamma)^\top \big(\mathbf{B}_{\gamma',+} + \mathbf{B}_{\gamma',-}\big)\mathbf{d}_b(\gamma).$$

By local stability, $a(\gamma') \leq \rho$, so (20) applied to γ' gives $\|\mathbf{B}_{\gamma',+} + \mathbf{B}_{\gamma',-}\|_{\mathrm{op}} \leq C(s_{\gamma'} + \varepsilon_n(t))$. Hence

$$H^{\mathrm{w}}(\gamma') \leq CQ^\circ + C\big(s_{\gamma'} + \varepsilon_n(t)\big) \sum_{b\in\{+,-\}} \|\mathbf{d}_b(\gamma)\|^2.$$

Applying (28) gives

$$H^{\mathrm{w}}(\gamma') \leq CQ^\circ + C\big(s_{\gamma'} + \varepsilon_n(t)\big)\big(s_\gamma + \varepsilon_n(t)\big)\big(H^{\mathrm{w}}(\gamma) + Q^\circ\big).$$

Since $a(\gamma), a(\gamma') \leq \rho$, we have $s_\gamma, s_{\gamma'} \leq r_0 + \rho$. Choose ρ, then r_0, and finally κ sufficiently small so that

$$C\big(s_{\gamma'} + \varepsilon_n(t)\big)\big(s_\gamma + \varepsilon_n(t)\big) \leq \frac{1}{2}.$$

Therefore,

$$H^{\mathrm{w}}(\gamma') \leq \frac{1}{2}H^{\mathrm{w}}(\gamma) + CQ^\circ,$$

which proves (27). □

D.3 Size of the oracle-instability term

We next bound Q° by treating separately the contributions from small realized margins, cross-group oracle prediction errors, and in-sample oracle prediction errors.

Lemma D.9. *With the constants from Lemma D.6, for every $t \geq 1$ satisfying $r \leq r_0$ and $p + t + \log(en) \leq \kappa n$, there exists an event $\mathcal{G}_{\mathrm{or}}(t)$ with probability at least $1 - Ce^{-t}$ such that*

$$\mathbb{E}\big[Q^{\circ}\mathbf{I}(\mathcal{G}_{\mathrm{loc}}(t) \cap \mathcal{G}_{\mathrm{or}}(t))\big] \leq Cr^3. \tag{31}$$

Moreover, with probability at least $1 - Ce^{-t}$,

$$Q^{\circ} \leq C\bigg(r^3 + r^{5/2}\sqrt{\frac{t}{n}} + r^2\frac{t}{n}\bigg). \tag{32}$$

When $r = 0$, $Q^{\circ} = 0$.

The proof is deferred to Appendix D.5.

D.4 Completion of local membership recovery

We now combine local stability, the contraction of the weighted Hamming loss, and the bound on Q° to prove Theorem 4.1.

Proof of Theorem 4.1. Let $\mathcal{E}_Q(t)$ be the event on which (32) in Lemma D.9 holds, and define

$$\mathcal{G}(t) := \mathcal{G}_{\mathrm{loc}}(t) \cap \mathcal{G}_{\mathrm{or}}(t) \cap \mathcal{E}_Q(t).$$

Then $\mathbb{P}(\mathcal{G}(t)^c) \leq Ce^{-t}$.

On $\mathcal{G}(t)$, Lemma D.7 gives $a(\gamma^{(k)}) \leq \rho$ for all $k \geq 0$, uniformly over starting pairs with $a(\gamma^{(0)}) \leq \rho/2$. Iterating Lemma D.8 therefore gives

$$H^{\mathrm{w}}(\gamma^{(k)}) \leq 2^{-k}H^{\mathrm{w}}(\gamma^{(0)}) + CQ^{\circ}\sum_{j=0}^{k-1} 2^{-j} \leq 2^{-k}H^{\mathrm{w}}(\gamma^{(0)}) + CQ^{\circ}.$$

Using (32) then yields

$$H^{\mathrm{w}}(\gamma^{(k)}) \leq 2^{-k}H^{\mathrm{w}}(\gamma^{(0)}) + C\bigg(r^3 + r^{5/2}\sqrt{\frac{t}{n}} + r^2\frac{t}{n}\bigg), \qquad k \geq 0,$$

which proves Theorem 4.1. □

Proof of Corollary 4.3. Let $\mathcal{G}(t)$ be the event constructed in the proof of Theorem 4.1.

Expectation bound. Let κ_0 be the constant in the dimension condition for $\mathcal{G}(t)$, and take $\kappa \leq \kappa_0/2$. Set $t_* := \kappa_0 n/2$ and $\mathcal{G}_* := \mathcal{G}(t_*)$. Since $p + \log(en) \leq \kappa n$,

$$p + t_* + \log(en) \leq \kappa_0 n, \qquad \mathbb{P}(\mathcal{G}_*^c) \leq Ce^{-\kappa_0 n/2}.$$

By the recursion in the proof of Theorem 4.1 and (31),

$$\mathbb{E}\big[H^{\mathrm{w}}(\gamma^{(K)})\mathbf{I}(\mathcal{G}_*)\big] \leq 2^{-K} + C\mathbb{E}\big[Q^{\circ}\mathbf{I}(\mathcal{G}_*)\big] \leq Cr^3,$$

where $\mathbb{E}H^{\mathrm{w}}(\gamma^{(0)}) \leq 1$ and $2^{-K} \leq r^3$.

Choosing $h = r$ in Fact D.4 and using the small-ball bound in Fact D.1, we obtain

$$\begin{aligned}\mathbb{E}H(\gamma^{(K)}) &\leq \mathbb{P}(|G| \leq r) + \frac{\mathbb{E}[H^{\mathrm{w}}(\gamma^{(K)})\mathbf{I}(\mathcal{G}_*)]}{r^2} + \mathbb{P}(\mathcal{G}_*^c) \\ &\leq Cr + Ce^{-\kappa_0 n/2}.\end{aligned}$$

Upper-tail bound. Fix $t \geq 1$ and set

$$W_t := r^3 + r^{5/2}\sqrt{\frac{t}{n}} + r^2\frac{t}{n}.$$

On $\mathcal{G}(t) \subseteq \mathcal{E}_{\mathrm{G}}(t)$, Corollary A.4 gives

$$H^{\mathrm{w}}(\gamma^{(0)}) \le \frac{1}{n}\sum_{i=1}^{n} G_i^2 = \mathbf{v}^\top \widehat{\mathbf{S}}\mathbf{v} \le \|\widehat{\mathbf{S}}\|_{\mathrm{op}} \le C.$$

Since $2^{-K} \le r^3 \le W_t$, Theorem 4.1 yields $H^{\mathrm{w}}(\gamma^{(K)}) \le CW_t$.

For deterministic $h > 0$, let

$$N_h := \sum_{i=1}^{n} \mathbf{I}(|G_i| \le h), \qquad p_h := \mathbb{P}(|G| \le h).$$

Then $N_h \sim \mathrm{Bin}(n, p_h)$ and $p_h \le h$ by Fact D.1. Bennett's inequality (Boucheron et al., 2013, Theorem 2.9) gives, with probability at least $1 - e^{-t}$,

$$\frac{N_h}{n} \le p_h + p_h b^{-1}\big(t/(np_h)\big),$$

where $b(u) := (1+u)\log(1+u) - u$. Using the bound $b^{-1}(v) \le C(\sqrt{v} + v/\log(e+v))$ for $v > 0$ (Wellner, 2017, Lemma 6.3), together with $p_h \le h$, gives

$$\begin{aligned}\frac{N_h}{n} &\le C\left(p_h + \sqrt{\frac{p_h t}{n}} + \frac{t/n}{\log(e + t/(np_h))}\right)\\ &\le C\left(h + \sqrt{\frac{ht}{n}} + \frac{t/n}{\log(e + t/(nh))}\right).\end{aligned}$$

Let $\mathcal{B}_h(t)$ denote the event on which this bound holds. On $\mathcal{G}(t) \cap \mathcal{B}_h(t)$, Fact D.4 gives

$$H(\gamma^{(K)}) \le \frac{N_h}{n} + \frac{CW_t}{h^2} \le C\left(h + \sqrt{\frac{ht}{n}} + \frac{t/n}{\log(e + t/(nh))} + \frac{W_t}{h^2}\right). \tag{33}$$

If $t/n \le r$, take $h = r$. Since $W_t/r^2 \le 3r$, (33) gives $H(\gamma^{(K)}) \le Cr$.

If $t/n > r$, set $L := \log(e + t/(nr))$ and $h := r\sqrt{L}$. Since $t/(nr) > 1$ and $\sup_{u \ge 1} \log^{5/2}(e+u)/u < \infty$, we have $L^{5/2} \le Ct/(nr)$. Together with $W_t \le 3r^2(t/n)$,

$$h \le C\frac{t/n}{L^2}, \qquad h + \sqrt{\frac{ht}{n}} + \frac{W_t}{h^2} \le C\frac{t/n}{L}.$$

Moreover,

$$\frac{t/n}{\log(e + t/(nh))} \le \frac{t/n}{L - \frac{1}{2}\log L} \le \frac{2(t/n)}{L}.$$

Hence (33) gives $H(\gamma^{(K)}) \le C(t/n)/\log(e + t/(nr))$.

Thus, with the corresponding deterministic choice of h in either regime, $H(\gamma^{(K)}) \le C\Psi_{n,r}(t)$ on $\mathcal{G}(t) \cap \mathcal{B}_h(t)$. Therefore

$$\mathbb{P}\big(H(\gamma^{(K)}) > C\Psi_{n,r}(t)\big) \le \mathbb{P}\big(\mathcal{G}(t)^c\big) + \mathbb{P}\big(\mathcal{B}_h(t)^c\big) \le Ce^{-t}.$$

Exact recovery. Since $H(\gamma^{(K)}) \in \{0, 1/n, \dots, 1\}$, $\mathbf{I}(H(\gamma^{(K)}) > 0) \le nH(\gamma^{(K)})$. Therefore

$$\mathbb{P}(H(\gamma^{(K)}) > 0) \le n\mathbb{E}H(\gamma^{(K)}) \le Cnr + Cne^{-\kappa_0 n/2} \le Cnr + Ce^{-\kappa_0 n/4}.$$

□

D.4.1 Finite-sample sharpness

Corollary D.10 (Matching tail lower bound)**.** *Let m and $B_{m,q(r)}$ be as in Proposition C.1. There are constants $r_*, \kappa_*, c, c', C > 0$, depending only on c_0, such that, for all sufficiently large n, if $0 < r \le r_*$ and $nr \ge 1$, then for every $1 \le t \le \kappa_* n$,*

$$\mathbb{P}(B_{m,q(r)} \ge cn\Psi_{n,r}(t)) \ge c'e^{-Ct}. \tag{34}$$

Consequently,

$$\inf_{\widehat{\mathbf{z}}} \sup_{\vartheta \in \Theta_{c_0}(n,p;\Lambda,\sigma)} \mathbb{P}_\vartheta(\bar{h}(\widehat{\mathbf{z}}, \mathbf{z}) \ge c\Psi_{n,r}(t)) \ge c'e^{-Ct}.$$

Proof. Write $B := B_{m,q(r)}$. Choose r_* so that $q(r) \le 1/4$ and $\kappa_* \le \zeta/2$. For all sufficiently large n, $1 \le t \le m$ and

$$\mathbb{E}B = mq(r) \asymp nr.$$

Choose $c > 0$ sufficiently small, with $c \le 1/6$.

If $t/n \le r$, then $n\Psi_{n,r}(t) \le 3nr$, so $cn\Psi_{n,r}(t) \le \mathbb{E}B/2$. Since $\mathbb{E}B^2 = (mq(r))^2 + mq(r)(1-q(r))$, the Paley–Zygmund inequality gives

$$\begin{aligned}\mathbb{P}\big(B \ge cn\Psi_{n,r}(t)\big) &\ge \mathbb{P}(B \ge \mathbb{E}B/2) \\ &\ge \frac{(\mathbb{E}B)^2}{4\mathbb{E}B^2} \ge \frac{mq(r)}{4(1+mq(r))} \ge c' > 0.\end{aligned}$$

If $t/n > r$, set $L := \log(e + t/(nr))$. As in the proof of Corollary 4.3,

$$n\Psi_{n,r}(t) = nr + \sqrt{nrt} + \frac{t}{L} \le C\frac{t}{L}.$$

Take $k := \lceil cn\Psi_{n,r}(t) \rceil$. Since $1 \le n\Psi_{n,r}(t) \le 3t$ and $t \le m$, we have $k \in [m]$ and $k \le C\frac{t}{L}$.

Using $mq(r) \asymp nr$,

$$\begin{aligned}k\log(e + k/(mq(r))) &\le C\frac{t}{L}\log(e + Ct/(nrL)) \\ &\le C\frac{t}{L}\big(L + \log C\big) \le Ct.\end{aligned}$$

Moreover, $\binom{m}{k} \ge (m/k)^k$ and $\log(1-q(r)) \ge -2q(r)$ give

$$\begin{aligned}\mathbb{P}(B = k) &= \binom{m}{k} q(r)^k (1-q(r))^{m-k} \\ &\ge \exp\big(-k\log(k/(mq(r))) - 2mq(r)\big) \\ &\ge \exp\big(-k\log(e + k/(mq(r))) - 2mq(r)\big) \ge e^{-Ct},\end{aligned}$$

where the last inequality also uses $mq(r) \le Cnr < Ct$. Thus (34) holds in both regimes.

Finally, Proposition C.1 with $k = \lceil cn\Psi_{n,r}(t) \rceil \le m$ gives

$$\begin{aligned}\inf_{\widehat{\mathbf{z}}} \sup_{\vartheta \in \Theta_{c_0}(n,p;\Lambda,\sigma)} \mathbb{P}_\vartheta\big(\bar{h}(\widehat{\mathbf{z}}, \mathbf{z}) \ge c\Psi_{n,r}(t)\big) &\ge \inf_{\widehat{\mathbf{z}}} \sup_{\vartheta \in \Theta_{c_0}(n,p;\Lambda,\sigma)} \mathbb{P}_\vartheta\big(n\bar{h}(\widehat{\mathbf{z}}, \mathbf{z}) \ge k\big) \\ &\ge \mathbb{P}(B \ge k) \ge c' e^{-Ct}.\end{aligned}$$

□

For $0 < r \le r_*$ with $nr \gtrsim 1$, Theorem 1.2 and Corollary D.10 give matching upper and lower Hamming-tail bounds at the scale $\Psi_{n,r}(t)$.

D.5 Auxiliary results for local membership recovery

This subsection collects the auxiliary bounds used above and concludes with the proof of Lemma D.9.

D.5.1 Gaussian wedges and selected second moments

For wedges of width s, the probability and selected second moments in arbitrary directions are of order s, while the selected second moment in the margin direction is of order s^2. The latter sharper bound is used in the local least-squares analysis. For a generic observation, we use the augmented-coordinate notation $\widetilde{\mathbf{u}}, \bar{\mathbf{v}}, G, \widetilde{\mathbf{u}}_\perp$.

Fix a sufficiently large constant C_0, depending only on c_0. For $s > 0$, define

$$\mathcal{W}(s) := \left\{ \big\{\widetilde{\mathbf{u}} : (G + \mathbf{h}_1^\top \widetilde{\mathbf{u}}_\perp)(G + \mathbf{h}_2^\top \widetilde{\mathbf{u}}_\perp) \le 0\big\} : \begin{array}{l} \mathbf{h}_1, \mathbf{h}_2 \perp \bar{\mathbf{v}}, \\ \|\mathbf{h}_1\| + \|\mathbf{h}_2\| \le C_0 s \end{array} \right\}.$$

The class $\mathcal{W}(s)$ has VC dimension $O(p)$, since its elements are fixed Boolean combinations of affine halfspaces in $\mathbb{R}^{p+1}$.

Lemma D.11 (Gaussian wedge geometry)**.** *For every $s > 0$ and every $E \in \mathcal{W}(s)$,*

$$\mathbb{P}(E) \le \frac{C_0}{\pi} s, \qquad \sup_{\substack{\mathbf{q}\in\mathbb{R}^p \\ \|\mathbf{q}\|=1}} \mathbb{E}\big[(\mathbf{q}^\top \mathbf{u})^2 \mathbf{I}(\widetilde{\mathbf{u}} \in E)\big] \le \frac{2C_0}{\pi} s. \tag{35}$$

Proof. For $E \in \mathcal{W}(s)$, let $\mathbf{h}_1, \mathbf{h}_2$ be its defining vectors and set

$$\mathbf{q}_j := \frac{\bar{\mathbf{v}} + \mathbf{h}_j}{\|\bar{\mathbf{v}} + \mathbf{h}_j\|}, \qquad j \in \{1, 2\}.$$

By (14) in Fact D.3, $E = \big\{(\mathbf{q}_1^\top \widetilde{\mathbf{u}})(\mathbf{q}_2^\top \widetilde{\mathbf{u}}) \le 0\big\}$. Hence, writing $\theta := \arccos(\mathbf{q}_1^\top \mathbf{q}_2)$, rotational invariance gives $\mathbb{P}(E) = \theta/\pi$.

Since $\|\bar{\mathbf{v}}\| = 1$ and $\mathbf{h}_j \perp \bar{\mathbf{v}}$, $\angle(\mathbf{q}_j, \bar{\mathbf{v}}) = \arctan \|\mathbf{h}_j\|$. Hence

$$\theta \le \sum_{j=1}^{2} \arctan \|\mathbf{h}_j\| \le \|\mathbf{h}_1\| + \|\mathbf{h}_2\| \le C_0 s.$$

Thus $\mathbb{P}(E) \le C_0 s/\pi$.

For the directional second moment, let $\mathcal{V} := \operatorname{span}(\mathbf{q}_1, \mathbf{q}_2)$ and decompose $(\mathbf{q}, 0) = \mathbf{q}_{\mathcal{V}} + \mathbf{q}_{\mathcal{V}^\perp}$. Since E depends only on $\widetilde{\mathbf{u}}_{\mathcal{V}}$, the independence of the Gaussian projections $\widetilde{\mathbf{u}}_{\mathcal{V}}$ and $\widetilde{\mathbf{u}}_{\mathcal{V}^\perp}$, together with $\mathbb{E}\widetilde{\mathbf{u}}_{\mathcal{V}^\perp} = 0$, gives

$$\mathbb{E}\big[(\mathbf{q}^\top \mathbf{u})^2 \mathbf{I}(E)\big] = \mathbb{E}\big[(\mathbf{q}_{\mathcal{V}}^\top \widetilde{\mathbf{u}}_{\mathcal{V}})^2 \mathbf{I}(E)\big] + \|\mathbf{q}_{\mathcal{V}^\perp}\|^2 \mathbb{P}(E).$$

Let $R := \|\widetilde{\mathbf{u}}_{\mathcal{V}}\|$. Since E is conic in $\mathcal{V}$, it depends only on the direction of $\widetilde{\mathbf{u}}_{\mathcal{V}}$, which is independent of R. Thus

$$\mathbb{E}\big[(\mathbf{q}_{\mathcal{V}}^\top \widetilde{\mathbf{u}}_{\mathcal{V}})^2 \mathbf{I}(E)\big] \le \|\mathbf{q}_{\mathcal{V}}\|^2 \mathbb{E}R^2\, \mathbb{P}(E) \le 2\|\mathbf{q}_{\mathcal{V}}\|^2 \mathbb{P}(E).$$

Therefore

$$\mathbb{E}\big[(\mathbf{q}^\top \mathbf{u})^2 \mathbf{I}(E)\big] \le \big(1 + \|\mathbf{q}_{\mathcal{V}}\|^2\big)\mathbb{P}(E) \le 2\mathbb{P}(E) \le \frac{2C_0}{\pi} s,$$

where $\mathbb{E}R^2 = \dim(\mathcal{V}) \le 2$ and $\|\mathbf{q}_{\mathcal{V}}\|^2 + \|\mathbf{q}_{\mathcal{V}^\perp}\|^2 = 1$. This proves (35).

□

Lemma D.12 (Uniform empirical moments over Gaussian wedges)**.** *Define*

$$\mathbf{B}_E := \frac{1}{n} \sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^\top \mathbf{I}(\widetilde{\mathbf{u}}_i \in E).$$

There exist constants $C, \psi_0, s_\star > 0$, depending only on C_0, such that for every $t \ge 1$ with $\psi_n(p,t) \le \psi_0$, with probability at least $1 - Ce^{-t}$, simultaneously for all $0 < s \le s_\star$ and $E \in \mathcal{W}(s)$,

$$\|\mathbf{B}_E\|_{\mathrm{op}} \le C\big(s + \varepsilon_n(t)\big), \qquad \mathbf{v}^\top \mathbf{B}_E \mathbf{v} = \frac{1}{n} \sum_{i=1}^{n} G_i^2 \mathbf{I}(\widetilde{\mathbf{u}}_i \in E) \le Cs^2. \tag{36}$$

Proof. We have

$$\|\mathbf{B}_E\|_{\mathrm{op}} = \sup_{\|\mathbf{q}\|=1} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{q}^\top \mathbf{u}_i)^2 \mathbf{I}(\widetilde{\mathbf{u}}_i \in E).$$

Take $\psi_0 < 1$ sufficiently small, set $L^2 := 2\log(2/\psi_n(p,t))$, and for $x \in [0, L^2]$, define

$$A_x := E \cap \{\sqrt{x} < |\mathbf{q}^\top \mathbf{u}| \le L\}.$$

As $E, \mathbf{q}, x$ vary, the sets A_x form a VC class of dimension $O(p)$. Hence Lemma A.1(ii) gives, uniformly in $x, E, \mathbf{q}$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(\widetilde{\mathbf{u}}_i \in A_x) \le \mathbb{P}(A_x) + C\sqrt{\mathbb{P}(A_x)\psi_n(p,t)} + C\psi_n(p,t). \tag{37}$$

Since

$$\int_0^{L^2} \mathbf{I}(\widetilde{\mathbf{u}} \in A_x)\,dx = (\mathbf{q}^\top \mathbf{u})^2 \mathbf{I}(|\mathbf{q}^\top \mathbf{u}| \le L)\mathbf{I}(\widetilde{\mathbf{u}} \in E),$$

integrating (37) over $x \in [0, L^2]$ gives

$$\frac{1}{n}\sum_{i=1}^n (\mathbf{q}^\top \mathbf{u}_i)^2 \mathbf{I}(|\mathbf{q}^\top \mathbf{u}_i| \le L)\mathbf{I}(\widetilde{\mathbf{u}}_i \in E) \le \int_0^{L^2} \mathbb{P}(A_x)\,dx + C\sqrt{\psi_n(p,t)} \int_0^{L^2} \sqrt{\mathbb{P}(A_x)}\,dx + CL^2\psi_n(p,t).$$

By Lemma D.11,

$$\int_0^{L^2} \mathbb{P}(A_x)\,dx = \mathbb{E}\big[(\mathbf{q}^\top \mathbf{u})^2 \mathbf{I}(|\mathbf{q}^\top \mathbf{u}| \le L)\mathbf{I}(E)\big] \le Cs.$$

Moreover, Cauchy–Schwarz gives

$$\int_0^{L^2} \sqrt{\mathbb{P}(A_x)}\,dx \le L\left(\int_0^{L^2} \mathbb{P}(A_x)\,dx\right)^{1/2} \le CL\sqrt{s}.$$

Hence

$$\frac{1}{n}\sum_{i=1}^n (\mathbf{q}^\top \mathbf{u}_i)^2 \mathbf{I}(|\mathbf{q}^\top \mathbf{u}_i| \le L)\mathbf{I}(\widetilde{\mathbf{u}}_i \in E) \le C\big(s + L\sqrt{s\psi_n(p,t)} + L^2\psi_n(p,t)\big).$$

Since $L\sqrt{s\psi_n(p,t)} \le \frac{1}{2}s + \frac{1}{2}L^2\psi_n(p,t)$ and $L^2\psi_n(p,t) \le C\varepsilon_n(t)$, we obtain

$$\frac{1}{n}\sum_{i=1}^n (\mathbf{q}^\top \mathbf{u}_i)^2 \mathbf{I}(|\mathbf{q}^\top \mathbf{u}_i| \le L)\mathbf{I}(\widetilde{\mathbf{u}}_i \in E) \le C\big(s + \varepsilon_n(t)\big). \tag{38}$$

For $J_{\mathbf{q},L} := \{i \in [n] : |\mathbf{q}^\top \mathbf{u}_i| > L\}$, the choice of L gives $\mathbb{P}(|\mathbf{q}^\top \mathbf{u}| > L) \le 2e^{-L^2/2} = \psi_n(p,t)$. Lemma A.1(ii), applied uniformly over $\mathbf{q}$, therefore gives $|J_{\mathbf{q},L}|/n \le C\psi_n(p,t)$. If $J_{\mathbf{q},L}$ is nonempty, Lemma A.2(ii) gives

$$\frac{1}{n}\sum_{i \in J_{\mathbf{q},L}} (\mathbf{q}^\top \mathbf{u}_i)^2 \le C\left(\frac{p+t}{n} + \frac{|J_{\mathbf{q},L}|}{n}\log\frac{en}{|J_{\mathbf{q},L}|}\right) \le C\varepsilon_n(t).$$

The same bound is immediate when $J_{\mathbf{q},L}$ is empty. Combining this with (38) gives

$$\|\mathbf{B}_E\|_{\mathrm{op}} \le C\big(s + \varepsilon_n(t)\big).$$

For the margin direction, let $\mathbf{h}_1, \mathbf{h}_2$ define E. Applying (15) with $\mathbf{a}_j = \bar{\mathbf{v}} + \mathbf{h}_j$ gives

$$\frac{1}{n}\sum_{i=1}^n G_i^2 \mathbf{I}(\widetilde{\mathbf{u}}_i \in E) \le \sum_{j=1}^2 \mathbf{h}_j^\top \left(\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i^\top\right)\mathbf{h}_j \le Cs^2\left\|\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i^\top\right\|_{\mathrm{op}} \le Cs^2,$$

where the last inequality follows from Lemma A.2(i). This proves the second bound in (36). □

D.5.2 Oracle and leave-one-out estimates

The oracle-instability part of Q° involves

$$\big(w_i + (Z_{ib}^\circ)^2\big)\mathbf{I}(|Z_{ib}^\circ| > \eta\sqrt{w_i}).$$

Since w_i depends only on $(\mathbf{u}_i, \xi_i)$, the treatment of Z°_{ib} differs according to whether observation i is used in constructing the oracle estimator. If $i \notin I^*_b$, $\mathbf{d}^\circ_b$ is independent of $(\mathbf{u}_i, \xi_i)$. For $i \in I^*_b$, we use a leave-one-out representation to separate the contribution of observation i.

Fix $b \in \{+, -\}$. For $i \in I^*_b$, define the leave-one-out counterparts of $\mathbf{S}^*_b$ and $\mathbf{d}^\circ_b$ as

$$\mathbf{S}^*_{b,-i} := \frac{1}{n} \sum_{\substack{j \in I^*_b \\ j \neq i}} \mathbf{u}_j \mathbf{u}_j^\top, \qquad \mathbf{d}^{\circ,-i}_b := r(\mathbf{S}^*_{b,-i})^{-1} \frac{1}{n} \sum_{\substack{j \in I^*_b \\ j \neq i}} \mathbf{u}_j \xi_j.$$

Then $\mathbf{d}^{\circ,-i}_b$ is independent of $(\mathbf{u}_i, \xi_i)$.

Let

$$\mathbf{P}_b := \mathbf{U}_{I^*_b} \big(\mathbf{U}_{I^*_b}^\top \mathbf{U}_{I^*_b}\big)^{-1} \mathbf{U}_{I^*_b}^\top = \frac{1}{n} \mathbf{U}_{I^*_b} (\mathbf{S}^*_b)^{-1} \mathbf{U}_{I^*_b}^\top,$$

be the least-squares projection matrix for group b. Write $P^{(b)}_{ij}$ for its (i, j) entry, $i, j \in I^*_b$, and define

$$h_{ib} := P^{(b)}_{ii} = \frac{1}{n} \mathbf{u}_i^\top (\mathbf{S}^*_b)^{-1} \mathbf{u}_i.$$

For $r > 0$, define

$$l_i := \frac{\mathbf{u}_i^\top \mathbf{d}^{\circ,-i}_b}{r}, \qquad i \in I^*_b.$$

The following are standard leave-one-out identities for least squares, stated here in the present notation.

Fact D.13. *Suppose $\mathbf{S}^*_b$ and $\mathbf{S}^*_{b,-i}$, $i \in I^*_b$, are invertible. Then $0 \le h_{ib} < 1$ and*

$$Z^\circ_{ib} = (1 - h_{ib}) \mathbf{u}_i^\top \mathbf{d}^{\circ,-i}_b + h_{ib} r \xi_i, \qquad i \in I^*_b. \tag{39}$$

For $r > 0$, conditional on $\mathbf{U}_{I^*_b}$,

$$l_i = \frac{1}{1 - h_{ib}} \sum_{\substack{j \in I^*_b \\ j \neq i}} P^{(b)}_{ij} \xi_j, \qquad l_i \sim N\big(0, h_{ib}/(1 - h_{ib})\big),$$

and l_i is independent of ξ_i.

Proof. The claims follow from the Sherman–Morrison formula applied to $\mathbf{S}^*_b = \mathbf{S}^*_{b,-i} + \mathbf{u}_i \mathbf{u}_i^\top / n$, together with the projector identity $\mathbf{P}_b^2 = \mathbf{P}_b$. □

Lemma D.14 (Uniform oracle error and leverage bounds)**.** *There exist constants $C, \kappa > 0$, depending only on c_0, such that for every $t \ge 1$ satisfying $p + t + \log(en) \le \kappa n$, there is an event $\mathcal{G}_{\mathrm{or}}(t) \subseteq \mathcal{E}_{\mathrm{G}}(t)$ of probability at least $1 - Ce^{-t}$ on which, simultaneously for both groups and every $i \in I^*_b$,*

$$\|\mathbf{d}^\circ_b\|, \|\mathbf{d}^{\circ,-i}_b\| \le Cr \sqrt{\frac{p + t + \log(en)}{n}}, \tag{40}$$

$$h_{ib} \le C \frac{p + t + \log(en)}{n}. \tag{41}$$

Proof. On $\mathcal{E}_{\mathrm{G}}(t)$, Corollary A.4 gives uniformly over b and $i \in I^*_b$,

$$\|(\mathbf{S}^*_b)^{-1}\|_{\mathrm{op}},\ \|(\mathbf{S}^*_{b,-i})^{-1}\|_{\mathrm{op}} \le C, \qquad \|\mathbf{u}_i\|^2 \le C\big(p + t + \log(en)\big).$$

Hence (41) follows directly.

By (17), conditional on the corresponding design,

$$\mathbf{d}^\circ_b \sim N\big(\mathbf{0}, \tfrac{r^2}{n}(\mathbf{S}^*_b)^{-1}\big), \qquad \mathbf{d}^{\circ,-i}_b \sim N\big(\mathbf{0}, \tfrac{r^2}{n}(\mathbf{S}^*_{b,-i})^{-1}\big).$$

Standard Gaussian norm bounds with confidence parameter $t + \log(en)$, followed by a union bound over b and i, give (40) with conditional failure probability at most Ce^{-t}. This gives the stated event $\mathcal{G}_{\mathrm{or}}(t) \subseteq \mathcal{E}_{\mathrm{G}}(t)$ with probability at least $1 - Ce^{-t}$. □

The following lemma controls the oracle-instability event when the coefficient error is independent of the target observation.

Lemma D.15 (Gaussian cone moment bound)**.** *Fix $\eta \in (0, 1/8)$ and $r > 0$. For a fixed $\mathbf{d} \in \mathbb{R}^p$, write*

$$\mathbf{d} = \zeta \mathbf{v} + \mathbf{d}_\perp, \qquad \mathbf{d}_\perp \perp \mathbf{v}.$$

If $\|\mathbf{d}\| \le \eta/4$, then for every integer $q \ge 1$,

$$\mathbb{E}\big[(w + (\mathbf{u}^\top \mathbf{d})^2)^q \mathbf{I}(|\mathbf{u}^\top \mathbf{d}| > \eta\sqrt{w})\big] \le q! C_\eta^q \frac{\|\mathbf{d}_\perp\|^{2q+2}}{r}.$$

Proof. Set $\tau := \|\mathbf{d}_\perp\|$ and $\mathcal{A} := \{|\mathbf{u}^\top \mathbf{d}| > \eta\sqrt{w}\}$. If $\tau = 0$, then $|\mathbf{u}^\top \mathbf{d}| = |\zeta G| \le (\eta/4)\sqrt{w}$, so $\mathcal{A}$ is empty. Assume $\tau > 0$ and write

$$\mathbf{u}^\top \mathbf{d} = \zeta G + \tau U, \qquad U \sim N(0, 1),$$

where U is independent of (G, ξ).

On $\mathcal{A}$, since $|\zeta| \le \eta/4$ and $|G| \le \sqrt{w}$,

$$\eta\sqrt{w} < |\zeta G + \tau U| \le \frac{\eta}{4}\sqrt{w} + \tau|U|,$$

and hence $\sqrt{w} \le C_\eta \tau |U|$. It follows that $|\mathbf{u}^\top \mathbf{d}| \le C_\eta \tau |U|$, so

$$w + (\mathbf{u}^\top \mathbf{d})^2 \le C_\eta \tau^2 U^2, \qquad \mathcal{A} \subseteq \{|G| \le C_\eta \tau |U|,\ |\xi| \le C_\eta \tau |U|/r\}.$$

Conditional on U,

$$\begin{aligned}\mathbb{P}(\mathcal{A} \mid U) &\le \mathbb{P}\left(|G| \le C_\eta \tau |U|,\ |\xi| \le C_\eta \tau |U|/r \mid U\right) \\ &= \mathbb{P}\left(|G| \le C_\eta \tau |U| \mid U\right) \mathbb{P}\left(|\xi| \le C_\eta \tau |U|/r \mid U\right) \\ &\le C_\eta \frac{\tau^2 U^2}{r},\end{aligned}$$

where the last inequality follows from the Gaussian small-ball bound. Therefore,

$$\mathbb{E}\big[(w + (\mathbf{u}^\top \mathbf{d})^2)^q \mathbf{I}(\mathcal{A}) \,\big|\, U\big] \le C_\eta^q \frac{\tau^{2q+2}|U|^{2q+2}}{r}.$$

Since $\mathbb{E}|U|^{2q+2} = (2q+1)!! \le 3^q q!$, taking expectations gives the claimed bound. □

D.5.3 In-sample oracle terms

The following lemma controls the in-sample oracle contribution to Q°.

Lemma D.16. *There exist constants $C, \kappa > 0$, depending only on c_0, such that if $r > 0$ and $t \ge 1$ satisfy $p + t + \log(en) \le \kappa n$, then with probability at least $1 - Ce^{-t}$, simultaneously for $b \in \{+, -\}$,*

$$\frac{1}{n} \sum_{i \in I_b^*} \big(w_i + (Z_{ib}^\circ)^2\big) \mathbf{I}(\mathcal{E}_{ib}^{\mathrm{or}}) \le C\left(r^3 + r^{5/2}\sqrt{\frac{t}{n}} + r^2 \frac{t}{n}\right).$$

Proof. Fix $b \in \{+, -\}$. Write $\boldsymbol{\xi}_b := (\xi_i)_{i \in I_b^*}$, $\mathbf{l}_b := (l_i)_{i \in I_b^*}$, and $\mathbf{D}_b := \mathrm{diag}(h_{ib} : i \in I_b^*)$. By Fact D.13,

$$\mathbf{l}_b = (\mathbf{I} - \mathbf{D}_b)^{-1}(\mathbf{P}_b - \mathbf{D}_b)\boldsymbol{\xi}_b =: \mathbf{L}_b \boldsymbol{\xi}_b.$$

On the design event given by (3) and (5), which has probability at least $1 - Ce^{-t}$, we have $\max_{i \in I_b^*} h_{ib} \le h_0$ for a fixed $0 < h_0 \le \min(\eta/2, 1/2)$, provided κ is sufficiently small. Hence $\mathrm{Var}(l_i \mid \mathbf{U}_{I_b^*}) = h_{ib}/(1 - h_{ib}) \le 2h_0$.

We first dominate the oracle contribution by a Lipschitz majorant. On $\mathcal{E}_{ib}^{\mathrm{or}}$, (39) gives

$$\eta\sqrt{w_i} < |Z_{ib}^\circ| = r|(1 - h_{ib})l_i + h_{ib}\xi_i| \le r|l_i| + h_0\sqrt{w_i}.$$

Thus $\sqrt{w_i} \le C_\eta r |l_i|$ and $w_i + (Z_{ib}^\circ)^2 \le C_\eta r^2 l_i^2$.

Choose $a_\eta > 0$ so that $a_\eta C_\eta \le 1/2$ and define

$$q_i(l,x) := \left(|l| - a_\eta\sqrt{\frac{G_i^2}{r^2} + x^2}\right)_+.$$

We have $q_i(l_i,\xi_i) \ge |l_i|/2$ on $\mathcal{E}_{ib}^{\mathrm{or}}$, and hence

$$\frac{1}{n}\sum_{i\in I_b^*}\big(w_i + (Z_{ib}^\circ)^2\big)\mathbf{I}(\mathcal{E}_{ib}^{\mathrm{or}}) \le \mathcal{R}_b := \frac{C_\eta r^2}{n}\sum_{i\in I_b^*} q_i(l_i,\xi_i)^2.$$

Moreover, $q_i(l,x)^2 \le l^2\mathbf{I}(|l| \ge a_\eta|G_i|/r)$.

We next bound the conditional mean of $\mathcal{R}_b$. Conditional on $\mathbf{U}_{I_b^*}$, l_i is Gaussian with variance at most $2h_0$. Hence the preceding bound and the truncated Gaussian second-moment inequality in Fact D.1 give

$$\mathbb{E}\big[q_i(l_i,\xi_i)^2 \mid \mathbf{U}_{I_b^*}\big] \le Ch_0\exp\left(-\frac{cG_i^2}{r^2h_0}\right),$$
$$\mathbb{E}\big[\mathcal{R}_b \mid \mathbf{U}_{I_b^*}\big] \le \frac{Cr^2h_0}{n}\sum_{i\in I_b^*}\exp\left(-\frac{cG_i^2}{r^2h_0}\right).$$

Let $Y_i := \exp(-cG_i^2/(r^2h_0))$. Then $0 \le Y_i \le 1$, $\mathbb{E}Y_i \le Cr\sqrt{h_0}$, and $\mathbb{E}Y_i^2 \le \mathbb{E}Y_i$. Bernstein's inequality gives, with probability at least $1-e^{-t}$,

$$\frac{1}{n}\sum_{i\in I_b^*} Y_i \le C\left(r + \sqrt{\frac{rt}{n}} + \frac{t}{n}\right).$$

On the intersection of these design events,

$$\mathbb{E}\big[\mathcal{R}_b \mid \mathbf{U}_{I_b^*}\big] \le C\left(r^3 + r^{5/2}\sqrt{\frac{t}{n}} + r^2\frac{t}{n}\right).$$

We finally control $\mathcal{R}_b$ by Gaussian concentration. Define

$$\mathbf{q}_b(\boldsymbol{\xi}_b) := \big(q_i((\mathbf{L}_b\boldsymbol{\xi}_b)_i,\xi_i)\big)_{i\in I_b^*}.$$

Since $x \mapsto \sqrt{G_i^2/r^2 + x^2}$ is 1-Lipschitz, $|q_i(l,x) - q_i(l',x')| \le |l-l'| + a_\eta|x-x'|$. By $\|\mathbf{L}_b\|_{\mathrm{op}} \le (1+h_0)/(1-h_0) \le 3$,

$$\|\mathbf{q}_b(\boldsymbol{\xi}_b) - \mathbf{q}_b(\boldsymbol{\xi}_b')\| \le C\|\boldsymbol{\xi}_b - \boldsymbol{\xi}_b'\|.$$

Therefore, by the definition of $\mathcal{R}_b$,

$$\big|\sqrt{\mathcal{R}_b(\boldsymbol{\xi}_b)} - \sqrt{\mathcal{R}_b(\boldsymbol{\xi}_b')}\big| \le \frac{Cr}{\sqrt{n}}\|\boldsymbol{\xi}_b - \boldsymbol{\xi}_b'\|.$$

Conditional on $\mathbf{U}_{I_b^*}$, $\boldsymbol{\xi}_b$ is standard Gaussian. Gaussian concentration for Lipschitz functions (Vershynin, 2018, Theorem 5.2.2) and Jensen's inequality give, with conditional probability at least $1-e^{-t}$,

$$\sqrt{\mathcal{R}_b} \le \sqrt{\mathbb{E}[\mathcal{R}_b \mid \mathbf{U}_{I_b^*}]} + Cr\sqrt{\frac{t}{n}},$$

and therefore

$$\mathcal{R}_b \le 2\mathbb{E}[\mathcal{R}_b \mid \mathbf{U}_{I_b^*}] + Cr^2\frac{t}{n}.$$

Combining the preceding bounds and taking a union bound over $b \in \{+,-\}$ proves the lemma. □

Proof of Lemma D.9. If $r = 0$, then $Z_{ib}^\circ = 0$ and $Q^\circ = 0$. Assume $r > 0$. For brevity, write

$$Q_i^{\mathrm{mar}} := w_i\mathbf{I}(\mathcal{E}_i^{\mathrm{mar}}), \qquad Q_{ib}^{\mathrm{or}} := \big(w_i + (Z_{ib}^\circ)^2\big)\mathbf{I}(\mathcal{E}_{ib}^{\mathrm{or}}),$$

so that $Q^\circ = n^{-1}\sum_i\big(Q_i^{\mathrm{mar}} + \sum_b Q_{ib}^{\mathrm{or}}\big)$.

Expectation bound. Conditioning on ξ and using the Gaussian small-ball bound gives $\mathbb{P}(|G| \le 8r|\xi| \mid \xi) \le Cr|\xi|$. Since $w \le 65r^2\xi^2$ on this event,

$$\mathbb{E}Q_i^{\mathrm{mar}} \le Cr^2\mathbb{E}\big[\xi^2\mathbb{P}(|G| \le 8r|\xi| \mid \xi)\big] \le Cr^3.$$

For sufficiently small κ, Lemma D.14 ensures on $\mathcal{G}_{\mathrm{or}}(t)$ that $\|\mathbf{d}_b^\circ\|, \|\mathbf{d}_b^{\circ,-i}\| \le \eta/8$ and $h_{ib} \le \eta/2$.

For $i \in I_{-b}^*$, condition on the sample in I_b^*. Since $\mathcal{G}_{\mathrm{loc}}(t) \cap \mathcal{G}_{\mathrm{or}}(t) \subseteq \{\|\mathbf{d}_b^\circ\| \le Cr\sqrt{(p+t+\log(en))/n}\}$, and the event on the right is measurable with respect to this sample, Lemma D.15 with $q = 1$ gives

$$\begin{aligned}
&\mathbb{E}\big[Q_{ib}^{\mathrm{or}}\mathbf{I}(\mathcal{G}_{\mathrm{loc}}(t) \cap \mathcal{G}_{\mathrm{or}}(t))\big] \\
&\quad\le \frac{C}{r}\mathbb{E}\Big[\|\mathbf{d}_{b,\perp}^\circ\|^4\mathbf{I}\Big(\|\mathbf{d}_b^\circ\| \le Cr\sqrt{(p+t+\log(en))/n}\Big)\Big] \\
&\quad\le \frac{C}{r}\left(Cr\sqrt{\frac{p+t+\log(en)}{n}}\right)^4 \le Cr^3.
\end{aligned}$$

For $i \in I_b^*$, (39) and $h_{ib} \le \eta/2$ imply

$$\mathcal{E}_{ib}^{\mathrm{or}} \subseteq \Big\{|\mathbf{u}_i^\top\mathbf{d}_b^{\circ,-i}| > \frac{\eta}{2}\sqrt{w_i}\Big\}, \qquad w_i + (Z_{ib}^\circ)^2 \le C\big(w_i + (\mathbf{u}_i^\top\mathbf{d}_b^{\circ,-i})^2\big).$$

Moreover, $\mathcal{G}_{\mathrm{loc}}(t) \cap \mathcal{G}_{\mathrm{or}}(t) \subseteq \{\|\mathbf{d}_b^{\circ,-i}\| \le Cr\sqrt{(p+t+\log(en))/n}\}$, where the event on the right is measurable with respect to the sample with observation i deleted. Conditioning on this sample and applying Lemma D.15 with cone parameter $\eta/2$ and $q = 1$ gives

$$\begin{aligned}
&\mathbb{E}\big[Q_{ib}^{\mathrm{or}}\mathbf{I}(\mathcal{G}_{\mathrm{loc}}(t) \cap \mathcal{G}_{\mathrm{or}}(t))\big] \\
&\quad\le \frac{C_\eta}{r}\mathbb{E}\Big[\|\mathbf{d}_{b,\perp}^{\circ,-i}\|^4\mathbf{I}\Big(\|\mathbf{d}_b^{\circ,-i}\| \le Cr\sqrt{(p+t+\log(en))/n}\Big)\Big] \\
&\quad\le \frac{C_\eta}{r}\left(Cr\sqrt{\frac{p+t+\log(en)}{n}}\right)^4 \le Cr^3.
\end{aligned}$$

Averaging over i, summing over b, and combining with the margin bound proves (31).

High-probability bound. The preceding calculation gives $\mathbb{E}Q_i^{\mathrm{mar}} \le Cr^3$. For every integer $q \ge 2$, the same conditioning argument as above, together with the Gaussian moment bound, gives

$$\mathbb{E}\big[(Q_i^{\mathrm{mar}})^q\big] \le C^q r^{2q+1}\mathbb{E}|\xi_i|^{2q+1} \le \frac{q!}{2}(Cr^5)(Cr^2)^{q-2}.$$

Thus the centered variables $Q_i^{\mathrm{mar}} - \mathbb{E}Q_i^{\mathrm{mar}}$ satisfy the Bernstein moment condition with variance proxy Cr^5 and scale Cr^2. Hence, with probability at least $1 - e^{-t}$,

$$\frac{1}{n}\sum_{i=1}^n Q_i^{\mathrm{mar}} \le C\left(r^3 + r^{5/2}\sqrt{\frac{t}{n}} + r^2\frac{t}{n}\right).$$

For the cross-group terms, fix b and condition on the sample in I_b^*. On the event $\|\mathbf{d}_b^\circ\| \le Cr\sqrt{(p+t+\log(en))/n} \le \eta/8$, the variables $(Q_{ib}^{\mathrm{or}})_{i\in I_{-b}^*}$ are conditionally independent, and Lemma D.15 gives a conditional mean of order $\|\mathbf{d}_{b,\perp}^\circ\|^4/r$ and a Bernstein moment condition with variance proxy $C\|\mathbf{d}_{b,\perp}^\circ\|^6/r$ and scale $C\|\mathbf{d}_{b,\perp}^\circ\|^2$. Therefore, with conditional probability at least $1 - e^{-t}$,

$$\frac{1}{n}\sum_{i\in I_{-b}^*} Q_{ib}^{\mathrm{or}} \le C\left(\frac{\|\mathbf{d}_{b,\perp}^\circ\|^4}{r} + \sqrt{\frac{\|\mathbf{d}_{b,\perp}^\circ\|^6 t}{rn}} + \|\mathbf{d}_{b,\perp}^\circ\|^2\frac{t}{n}\right).$$

Using (40) on $\mathcal{G}_{\mathrm{or}}(t)$ bounds the right-hand side by $C(r^3 + r^{5/2}\sqrt{t/n} + r^2t/n)$. Thus the cross-group bounds hold simultaneously for both groups with failure probability at most Ce^{-t}.

Lemma D.16 gives the same bound for the in-sample oracle terms. Combining the margin, cross-group, and in-sample bounds proves (32). □

E INITIALIZATION AND COMPLETION OF EFFICIENT RECOVERY

This appendix analyzes the initialization stage of Algorithm 3.1 and completes the proofs of Theorem 1.2 and Corollary 1.3. We first control the pooled estimator and residual scale. We then analyze bounded residual weighting, identify its population spectral signal, and establish a uniform empirical approximation. The resulting spectral direction, together with the pooled estimator, yields a first least-squares refit in the local neighborhood required by Theorem 4.1. Combining this initialization with the local contraction result gives the efficient-recovery guarantees, including the noiseless exact-recovery result. The technical proof of the uniform spectral approximation is deferred to the final subsection.

We use the whitening notation of Appendix A. Recall from Section 5 that

$$\omega := \frac{|I_+^*|}{n} \in [c_0, 1-c_0], \qquad \bar{\boldsymbol{\beta}} := \omega \boldsymbol{\beta}_+^* + (1-\omega)\boldsymbol{\beta}_-^*.$$

E.1 POOLED LEAST SQUARES AND RESIDUAL SCALE

We first control the pooled least-squares estimator and its residual scale. Write

$$\widehat{\boldsymbol{\alpha}} := \mathbf{A}\widehat{\mathbf{b}}, \qquad \boldsymbol{\delta}_i := \boldsymbol{\alpha}_{z_i^*}^* - \bar{\boldsymbol{\alpha}} = \begin{cases} (1-\omega)\Lambda \mathbf{v}, & z_i^* = +1, \\ -\omega \Lambda \mathbf{v}, & z_i^* = -1, \end{cases}$$

and define the residual at the population pooled coefficient by

$$\eta_i := y_i - \mathbf{u}_i^\top \bar{\boldsymbol{\alpha}} = \mathbf{u}_i^\top \boldsymbol{\delta}_i + \epsilon_i.$$

Let $\boldsymbol{\eta} := (\eta_i)_{i=1}^n$, so that $\mathbf{y} = \mathbf{U}\bar{\boldsymbol{\alpha}} + \boldsymbol{\eta}$. The following identities will be used to control $\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}$ and $\widehat{T}$.

The initialization analysis requires $\widehat{\mathbf{b}}$ concentrating around the population pooled coefficient and $\widehat{T}$ being comparable to Λ.

Fact E.1. *The pooled least-squares estimator satisfies*

$$\mathbf{U}^\top \mathbf{U}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}) = \sum_{i=1}^n (\mathbf{u}_i \mathbf{u}_i^\top - \mathbf{I}_p)\boldsymbol{\delta}_i + \sum_{i=1}^n \mathbf{u}_i \epsilon_i. \tag{42}$$

Moreover,

$$\widehat{T}^2 = \frac{1}{n}\|\boldsymbol{\eta}\|^2 - \frac{1}{n}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}})^\top \mathbf{U}^\top \mathbf{U}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}). \tag{43}$$

Proof. The pooled normal equation gives

$$\mathbf{U}^\top \boldsymbol{\eta} = \mathbf{U}^\top \mathbf{U}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}).$$

Since $\sum_i \boldsymbol{\delta}_i = \mathbf{0}$, $\sum_i \mathbf{u}_i \mathbf{u}_i^\top \boldsymbol{\delta}_i = \sum_i (\mathbf{u}_i \mathbf{u}_i^\top - \mathbf{I}_p)\boldsymbol{\delta}_i$, so (42) follows. Moreover,

$$n\widehat{T}^2 = \|\boldsymbol{\eta} - \mathbf{U}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}})\|^2 = \|\boldsymbol{\eta}\|^2 - (\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}})^\top \mathbf{U}^\top \mathbf{U}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}).$$

This proves (43). □

Lemma E.2. *There exist constants $C, \kappa, r_0 > 0$, depending only on c_0, such that for every $t \geq 1$ satisfying $p + t \leq \kappa n$, with probability at least $1 - Ce^{-t}$,*

$$\frac{\|\widehat{\mathbf{b}} - \bar{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}}}{\Lambda} \leq C(1+r)\sqrt{\frac{p+t}{n}}. \tag{44}$$

If in addition $r \leq r_0$, then on the same event

$$C^{-1}\Lambda \leq \widehat{T} \leq C\Lambda. \tag{45}$$

Proof. By duality,

$$\left\|\mathbf{U}^\top \mathbf{U}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}})\right\| = \sup_{\|\mathbf{q}\|=1} \left|\mathbf{q}^\top \mathbf{U}^\top \mathbf{U}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}})\right|.$$

For every fixed unit vector $\mathbf{q}$, (42) gives

$$\mathbf{q}^\top \mathbf{U}^\top \mathbf{U}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}) = \sum_{i=1}^{n} \big[\mathbf{q}^\top (\mathbf{u}_i \mathbf{u}_i^\top - \mathbf{I}_p)\boldsymbol{\delta}_i + (\mathbf{q}^\top \mathbf{u}_i)\epsilon_i\big].$$

The summands are independent, centered, and sub-exponential, with sub-exponential norms bounded by $C(\Lambda + \sigma)$. Bernstein's inequality and a standard sphere-covering argument (see, e.g., Vershynin, 2018, Corollary 4.2.13 and Exercise 4.4.2) therefore give, with probability at least $1 - Ce^{-t}$,

$$\frac{1}{n}\big\|\mathbf{U}^\top \mathbf{U}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}})\big\| \le C(\Lambda + \sigma)\left(\sqrt{\frac{p+t}{n}} + \frac{p+t}{n}\right).$$

On the event in Lemma A.2(i), $n^{-1}\mathbf{U}^\top \mathbf{U}$ is uniformly well conditioned. Hence, for sufficiently small κ,

$$\|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\| \le C\Lambda(1+r)\sqrt{\frac{p+t}{n}}.$$

Since $\|\widehat{\mathbf{b}} - \bar{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}} = \|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\|$, this proves (44).

For the scale bound, the η_i are independent centered Gaussian variables with

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\eta_i^2 = \Lambda^2\omega(1-\omega) + \sigma^2 \asymp \Lambda^2.$$

For $r \le r_0$, the centered variables $\eta_i^2 - \mathbb{E}\eta_i^2$ are sub-exponential with ψ_1-norm at most $C\Lambda^2$. Hence Bernstein's inequality gives, with probability at least $1 - Ce^{-t}$,

$$\frac{1}{n}\|\boldsymbol{\eta}\|^2 \asymp \Lambda^2\omega(1-\omega) + \sigma^2 \asymp \Lambda^2,$$

for $t \le \kappa n$.

By (43), (44), and Lemma A.2(i),

$$\widehat{T}^2 = \frac{1}{n}\|\boldsymbol{\eta}\|^2 + O\left(\Lambda^2(1+r)^2\frac{p+t}{n}\right) \asymp \Lambda^2$$

for sufficiently small κ and r_0. This proves (45). □

E.2 Residual weighting and population spectral signal

Finite-sample limitation of squared-residual weighting. As shown in Section 5, squared-residual weighting has the correct population contrast direction. Its finite-sample limitation is captured by the following proposition.

Proposition E.3. *There exists a universal constant $C > 1$ such that, for all sufficiently large n, if $p \ge C\log n$, then with probability at least C^{-1},*

$$\Big\|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2(\mathbf{u}_i\mathbf{u}_i^\top - \mathbf{I}_p)\Big\|_{\mathrm{op}} \ge C^{-1}\sigma^2\frac{p\log n}{n} - C\sigma^2.$$

Proof. By standard Gaussian and chi-square tail bounds, there exist constants $a, b, c > 0$ and $\alpha < 1$ such that

$$\mathbb{P}(\epsilon_i^2 \ge a\sigma^2\log n) \ge n^{-\alpha}, \qquad \mathbb{P}(\|\mathbf{u}_i\|^2 \ge bp) \ge c.$$

By independence,

$$\mathbb{P}\Big(\exists j \in [n] : \epsilon_j^2 \ge a\sigma^2\log n,\ \|\mathbf{u}_j\|^2 \ge bp\Big) \ge 1 - (1 - cn^{-\alpha})^n \ge c,$$

for all sufficiently large n. Hence, with probability at least c, there exists $j \in [n]$ such that $\epsilon_j^2 \ge a\sigma^2\log n$ and $\|\mathbf{u}_j\|^2 \ge bp$.

For each j, let $\mathbf{q}_j := \mathbf{u}_j/\|\mathbf{u}_j\|$. Then

$$\mathbf{q}_j^\top \Big[\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\mathbf{u}_i\mathbf{u}_i^\top - \mathbf{I}_p)\Big]\mathbf{q}_j = \frac{\epsilon_j^2}{n}(\|\mathbf{u}_j\|^2 - 1) + R_j,$$

where $R_j := \frac{1}{n}\sum_{i\neq j}\epsilon_i^2\big((\mathbf{q}_j^\top \mathbf{u}_i)^2 - 1\big)$. Conditional on $\mathbf{q}_j$, R_j is centered and $\mathrm{Var}(R_j \mid \mathbf{q}_j) \le C\sigma^4/n$. Hence Chebyshev's inequality and a union bound give

$$\mathbb{P}\Big(\min_{j\in[n]} R_j < -A\sigma^2\Big) \le \frac{C}{A^2}.$$

Choose A sufficiently large so that $C/A^2 \le c/2$. Combining this bound with the preceding large-deviation event, with probability at least $c/2$, there exists $j \in [n]$ such that

$$\epsilon_j^2 \ge a\sigma^2 \log n, \qquad \|\mathbf{u}_j\|^2 \ge bp, \qquad R_j \ge -A\sigma^2.$$

For such j,

$$\mathbf{q}_j^\top \Big[\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\mathbf{u}_i\mathbf{u}_i^\top - \mathbf{I}_p)\Big]\mathbf{q}_j \ge c\sigma^2\frac{p\log n}{n} - C\sigma^2.$$

This yields the lower bound for the operator norm and proves the claim. □

The proposition shows that squared-residual weighting introduces a logarithmic loss in spectral concentration. A similar phenomenon appears in spectral initialization for phase retrieval, where truncation is used to achieve optimal sample complexity; see, for example, Chen & Candès (2017). Our bounded transform is automatically scaled by the pooled residual scale $\widehat{T}$ and requires no additional tuning parameter. We next verify that it preserves the same population contrast direction.

Bounded residual weighting. Recall $h_T(x) = 1 - e^{-x^2/(2T^2)}$. We first give some simple facts about h_T to control perturbations in the coefficient and scale parameters.

Fact E.4. *For every $T > 0$ and $x, y \in \mathbb{R}$,*

$$h_T'(x) = \frac{x}{T^2}e^{-x^2/(2T^2)}, \qquad h_T''(x) = \left(\frac{1}{T^2} - \frac{x^2}{T^4}\right)e^{-x^2/(2T^2)},$$

and

$$|h_T(x) - h_T(y)| \le C\min(1, |x-y|/T), \qquad |\partial_{\log T} h_T(x)| \le 2/e.$$

If $Q \sim N(0, s^2)$ with $s \ge 0$, then

$$\mathbb{E}h_T''(Q) = \frac{T}{(T^2+s^2)^{3/2}}.$$

Proof. Direct differentiation gives the derivative formulas, including $\partial_{\log T} h_T(x) = -(x/T)^2 e^{-x^2/(2T^2)}$. The difference and scale-derivative bounds follow from $0 \le h_T \le 1$, $\sup_x |h_T'(x)| \le C/T$, and $\sup_{u\ge 0} ue^{-u/2} = 2/e$. Finally,

$$\mathbb{E}e^{-Q^2/(2T^2)} = \frac{T}{\sqrt{T^2+s^2}}, \qquad \mathbb{E}\big[Q^2 e^{-Q^2/(2T^2)}\big] = \frac{s^2T^3}{(T^2+s^2)^{3/2}},$$

and substitution into the expression for h_T'' gives the last identity. □

Fact E.5 (Gaussian identity for bounded residual weights)**.** *For any fixed $\mathbf{d} \in \mathbb{R}^p$ and $T > 0$,*

$$\mathbb{E}\big[h_T(\mathbf{u}^\top\mathbf{d} + \epsilon)(\mathbf{u}\mathbf{u}^\top - \mathbf{I}_p)\big] = \frac{T}{(T^2 + \|\mathbf{d}\|^2 + \sigma^2)^{3/2}}\mathbf{d}\mathbf{d}^\top.$$

Proof. Conditional on ϵ, the second-order Gaussian Stein identity applied to $h_T(\mathbf{u}^\top\mathbf{d} + \epsilon)$ gives

$$\mathbb{E}_{\mathbf{u}}\big[h_T(\mathbf{u}^\top\mathbf{d} + \epsilon)(\mathbf{u}\mathbf{u}^\top - \mathbf{I}_p) \mid \epsilon\big] = \mathbb{E}_{\mathbf{u}}\big[h_T''(\mathbf{u}^\top\mathbf{d} + \epsilon) \mid \epsilon\big]\mathbf{d}\mathbf{d}^\top.$$

Taking expectation over ϵ,

$$\mathbb{E}\big[h_T(\mathbf{u}^\top \mathbf{d} + \epsilon)(\mathbf{u}\mathbf{u}^\top - \mathbf{I}_p)\big] = \mathbb{E}\big[h_T''(\mathbf{u}^\top \mathbf{d} + \epsilon)\big]\mathbf{d}\mathbf{d}^\top.$$

Since $\mathbf{u}^\top \mathbf{d} + \epsilon \sim N(0, \|\mathbf{d}\|^2 + \sigma^2)$, Fact E.4 yields

$$\mathbb{E}\big[h_T''(\mathbf{u}^\top \mathbf{d} + \epsilon)\big] = \frac{T}{(T^2 + \|\mathbf{d}\|^2 + \sigma^2)^{3/2}}.$$

Substitution gives the claimed identity. □

Population spectral structure. We now apply Fact E.5 to the two componentwise residuals. For a deterministic coefficient $\mathbf{b}$, their slopes in whitened coordinates are

$$\boldsymbol{\delta}_+(\mathbf{b}) := (1-\omega)\Lambda\mathbf{v} - \mathbf{A}(\mathbf{b} - \bar{\boldsymbol{\beta}}), \qquad \boldsymbol{\delta}_-(\mathbf{b}) := -\omega\Lambda\mathbf{v} - \mathbf{A}(\mathbf{b} - \bar{\boldsymbol{\beta}}).$$

For deterministic $(\boldsymbol{\alpha}, T)$, define the population spectral operator

$$\mathbf{M}_U(\boldsymbol{\alpha}, T) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[h_T(y_i - \mathbf{u}_i^\top \boldsymbol{\alpha})(\mathbf{u}_i\mathbf{u}_i^\top - \mathbf{I}_p)\big].$$

For $T > 0$ and $\mathbf{d} \in \mathbb{R}^p$, define

$$F_T(\mathbf{d}) := \frac{T\mathbf{d}\mathbf{d}^\top}{(T^2 + \|\mathbf{d}\|^2 + \sigma^2)^{3/2}}.$$

Fact E.6 (Population spectral representation)**.** *For every deterministic $\mathbf{b} \in \mathbb{R}^p$ and $T > 0$,*

$$\mathbf{M}_U(\mathbf{A}\mathbf{b}, T) = \omega F_T(\boldsymbol{\delta}_+(\mathbf{b})) + (1-\omega)F_T(\boldsymbol{\delta}_-(\mathbf{b})). \tag{46}$$

Moreover,

$$\mathbf{M}_U(\bar{\boldsymbol{\alpha}}, T) = \lambda(T)\mathbf{v}\mathbf{v}^\top, \tag{47}$$

where

$$\lambda(T) = \omega\frac{T(1-\omega)^2\Lambda^2}{(T^2 + (1-\omega)^2\Lambda^2 + \sigma^2)^{3/2}} + (1-\omega)\frac{T\omega^2\Lambda^2}{(T^2 + \omega^2\Lambda^2 + \sigma^2)^{3/2}}.$$

For fixed $0 < A_- < A_+$, if $A_-\Lambda \le T \le A_+\Lambda$ and $r \le r_0$, then

$$c \le \lambda(T) \le C, \tag{48}$$

where $c, C > 0$ depend only on c_0, A_-, A_+, and r_0.

Proof. Apply Fact E.5 separately to the two groups in whitened coordinates to obtain the representation. At $\mathbf{b} = \bar{\boldsymbol{\beta}}$,

$$\boldsymbol{\delta}_+(\bar{\boldsymbol{\beta}}) = (1-\omega)\Lambda\mathbf{v}, \qquad \boldsymbol{\delta}_-(\bar{\boldsymbol{\beta}}) = -\omega\Lambda\mathbf{v}.$$

Substitution gives (47) and the stated expression for $\lambda(T)$. After normalization by Λ, $\lambda(T)$ is a positive continuous function of $(T/\Lambda, \omega, r)$ on $[A_-, A_+] \times [c_0, 1-c_0] \times [0, r_0]$. The bounds in (48) therefore follow. □

Fact E.7 (Population spectral perturbation)**.** *Fix $0 < A_- < A_+$. There exists $C > 0$, depending only on A_-, A_+, such that, for all $\boldsymbol{\alpha}, \boldsymbol{\alpha}_0 \in \mathbb{R}^p$ and $A_-\Lambda \le T, T_0 \le A_+\Lambda$,*

$$\|\mathbf{M}_U(\boldsymbol{\alpha}, T)\|_{\mathrm{op}} \le C, \tag{49}$$

and

$$\|\mathbf{M}_U(\boldsymbol{\alpha}, T) - \mathbf{M}_U(\boldsymbol{\alpha}_0, T_0)\|_{\mathrm{op}} \le C\left(\frac{\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|}{\Lambda} + |\log(T/T_0)|\right). \tag{50}$$

In particular, for every $\mathbf{b} \in \mathbb{R}^p$ and $A_-\Lambda \le T \le A_+\Lambda$,

$$\|\mathbf{M}_U(\mathbf{A}\mathbf{b}, T) - \lambda(T)\mathbf{v}\mathbf{v}^\top\|_{\mathrm{op}} \le C\frac{\|\mathbf{b} - \bar{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}}}{\Lambda}. \tag{51}$$

Proof. Write $s_d^2 := T^2 + \|\mathbf{d}\|^2 + \sigma^2$. Direct differentiation gives

$$DF_T(\mathbf{d})[\mathbf{h}] = \frac{T(\mathbf{h}\mathbf{d}^\top + \mathbf{d}\mathbf{h}^\top)}{s_d^3} - \frac{3T(\mathbf{d}^\top\mathbf{h})\mathbf{d}\mathbf{d}^\top}{s_d^5},$$

and

$$\partial_{\log T} F_T(\mathbf{d}) = \frac{T\mathbf{d}\mathbf{d}^\top}{s_d^3} - \frac{3T^3\mathbf{d}\mathbf{d}^\top}{s_d^5}.$$

Since $T \le s_d$ and $\|\mathbf{d}\| \le s_d$, the mean-value theorem yields

$$\|F_T(\mathbf{d})\|_{\mathrm{op}} \le 1, \qquad \|DF_T(\mathbf{d})[\mathbf{h}]\|_{\mathrm{op}} \le C\frac{\|\mathbf{h}\|}{T}, \qquad \|\partial_{\log T} F_T(\mathbf{d})\|_{\mathrm{op}} \le C.$$

Therefore, for $A_-\Lambda \le T, T_0 \le A_+\Lambda$,

$$\|F_T(\mathbf{d}) - F_T(\mathbf{d}_0)\|_{\mathrm{op}} \le C\frac{\|\mathbf{d} - \mathbf{d}_0\|}{\Lambda}, \qquad \|F_T(\mathbf{d}) - F_{T_0}(\mathbf{d})\|_{\mathrm{op}} \le C|\log(T/T_0)|.$$

Now write $\boldsymbol{\alpha} = \mathbf{A}\mathbf{b}$ and $\boldsymbol{\alpha}_0 = \mathbf{A}\mathbf{b}_0$. Since $\boldsymbol{\delta}_\pm(\mathbf{b}) - \boldsymbol{\delta}_\pm(\mathbf{b}_0) = -(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)$, the representation (46) immediately gives (49) and (50).

Finally, taking $\boldsymbol{\alpha}_0 = \bar{\boldsymbol{\alpha}}$ and $T_0 = T$, and using (47), gives

$$\|\mathbf{M}_U(\mathbf{A}\mathbf{b}, T) - \lambda(T)\mathbf{v}\mathbf{v}^\top\|_{\mathrm{op}} \le C\frac{\|\mathbf{A}(\mathbf{b} - \bar{\boldsymbol{\beta}})\|}{\Lambda},$$

which proves (51). □

E.3 Uniform empirical spectral approximation

By Lemma E.2, $(\widehat{\boldsymbol{\alpha}}, \widehat{T})$ lies in a fixed neighborhood of $(\bar{\boldsymbol{\alpha}}, \Lambda)$ with high probability. We handle its data dependence by establishing a concentration bound uniformly over this neighborhood.

For deterministic $(\boldsymbol{\alpha}, T)$, define

$$\widetilde{\mathbf{M}}_U(\boldsymbol{\alpha}, T) := \frac{1}{n}\sum_{i=1}^n h_T(y_i - \mathbf{u}_i^\top\boldsymbol{\alpha})(\mathbf{u}_i\mathbf{u}_i^\top - \widehat{\mathbf{S}}).$$

At $(\widehat{\boldsymbol{\alpha}}, \widehat{T})$, the spectral matrix in Algorithm 3.1 can be written as

$$\widehat{\mathbf{M}} = \mathbf{A}\widetilde{\mathbf{M}}_U(\widehat{\boldsymbol{\alpha}}, \widehat{T})\mathbf{A}. \tag{52}$$

We first establish uniform concentration with the population centering $\mathbf{I}_p$. The proof is deferred to Appendix E.6.

Lemma E.8. *Fix $\delta, \eta > 0$ and $0 < A_- < A_+$. There exists $\kappa > 0$, depending only on these constants and c_0, such that, for every $t \ge 1$ satisfying $p + t + \log(en) \le \kappa n$, with probability at least $1 - Ce^{-t}$,*

$$\sup_{\substack{\|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\| \le \delta\Lambda \\ A_-\Lambda \le T \le A_+\Lambda}} \Big\|\frac{1}{n}\sum_{i=1}^n h_T(y_i - \mathbf{u}_i^\top\boldsymbol{\alpha})(\mathbf{u}_i\mathbf{u}_i^\top - \mathbf{I}_p) - \mathbf{M}_U(\boldsymbol{\alpha}, T)\Big\|_{\mathrm{op}} \le \eta.$$

We next account for the empirical centering by $\widehat{\mathbf{S}}$ and the whitening by $\widehat{\mathbf{S}}^{-1/2}$ arising from the generalized eigenproblem.

Lemma E.9. *Fix $\delta, \eta > 0$ and $0 < A_- < A_+$. There exists $\kappa > 0$, depending only on these constants and c_0, such that, for every $t \ge 1$ satisfying $p + t + \log(en) \le \kappa n$, with probability at least $1 - Ce^{-t}$,*

$$\sup_{\substack{\|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\| \le \delta\Lambda \\ A_-\Lambda \le T \le A_+\Lambda}} \Big\|\widehat{\mathbf{S}}^{-1/2}\widetilde{\mathbf{M}}_U(\boldsymbol{\alpha}, T)\widehat{\mathbf{S}}^{-1/2} - \mathbf{M}_U(\boldsymbol{\alpha}, T)\Big\|_{\mathrm{op}} \le \eta.$$

Proof. Apply Lemma E.8 with tolerance $\eta_0 > 0$ and intersect its event with $\mathcal{E}_{\mathrm{G}}(t)$ from Corollary A.4. Then

$$\|\widehat{\mathbf{S}} - \mathbf{I}_p\|_{\mathrm{op}} \le C\sqrt{\frac{p+t}{n}}.$$

By definition,

$$\widetilde{\mathbf{M}}_U(\boldsymbol{\alpha}, T) = \frac{1}{n}\sum_{i=1}^{n} h_T(y_i - \mathbf{u}_i^\top \boldsymbol{\alpha})(\mathbf{u}_i\mathbf{u}_i^\top - \mathbf{I}_p) - \left(\frac{1}{n}\sum_{i=1}^{n} h_T(y_i - \mathbf{u}_i^\top \boldsymbol{\alpha})\right)(\widehat{\mathbf{S}} - \mathbf{I}_p).$$

Since $0 \le h_T \le 1$, uniformly over the parameter neighborhood,

$$\|\widetilde{\mathbf{M}}_U(\boldsymbol{\alpha}, T) - \mathbf{M}_U(\boldsymbol{\alpha}, T)\|_{\mathrm{op}} \le \eta_0 + C\sqrt{\frac{p+t}{n}}.$$

For sufficiently small κ,

$$\|\widehat{\mathbf{S}}^{-1/2}\|_{\mathrm{op}} \le C, \qquad \|\widehat{\mathbf{S}}^{-1/2} - \mathbf{I}_p\|_{\mathrm{op}} \le C\sqrt{\frac{p+t}{n}}.$$

Moreover, by (49) in Fact E.7, $\|\mathbf{M}_U(\boldsymbol{\alpha}, T)\|_{\mathrm{op}} \le C$ uniformly on the same neighborhood. Hence

$$\begin{aligned}
&\big\|\widehat{\mathbf{S}}^{-1/2}\widetilde{\mathbf{M}}_U(\boldsymbol{\alpha}, T)\widehat{\mathbf{S}}^{-1/2} - \mathbf{M}_U(\boldsymbol{\alpha}, T)\big\|_{\mathrm{op}} \\
&\quad \le C\|\widetilde{\mathbf{M}}_U(\boldsymbol{\alpha}, T) - \mathbf{M}_U(\boldsymbol{\alpha}, T)\|_{\mathrm{op}} + C\|\widehat{\mathbf{S}}^{-1/2} - \mathbf{I}_p\|_{\mathrm{op}} \\
&\quad \le C\eta_0 + C\sqrt{\frac{p+t}{n}}.
\end{aligned}$$

Choosing η_0 and then κ sufficiently small proves the claim. □

E.4 Spectral direction and first refit

We first show that, starting from a coefficient close to $\bar{\boldsymbol{\beta}}$ and a direction close to $\boldsymbol{\theta}$, one assignment step followed by least-squares refitting produces a coefficient pair in the local neighborhood. We then apply this result to $(\widehat{\mathbf{b}}, \widehat{\mathbf{w}})$.

For $(\mathbf{b}, \mathbf{w})$, let $\Gamma(\mathbf{b}, \mathbf{w})$ denote the coefficient pair obtained by assigning according to the sign of $(\mathbf{x}_i^\top \mathbf{w})(y_i - \mathbf{x}_i^\top \mathbf{b})$ and refitting the two assigned groups by least squares.

Lemma E.10. *For the radius ρ in Theorem 4.1, there exist $\delta_{\mathrm{ref}}, r_0, \kappa > 0$, depending only on c_0, such that, for every $t \ge 1$ satisfying $p + t + \log(en) \le \kappa n$ and $r \le r_0$, with probability at least $1 - Ce^{-t}$, simultaneously for all $(\mathbf{b}, \mathbf{w})$ satisfying*

$$\frac{\|\mathbf{b} - \bar{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}}}{\Lambda} \le \delta_{\mathrm{ref}}, \qquad \sin\angle(\mathbf{A}\mathbf{w}, \mathbf{A}\boldsymbol{\theta}) \le \delta_{\mathrm{ref}},$$

we have

$$a(\Gamma(\mathbf{b}, \mathbf{w})) \le \rho/2.$$

Proof. Fix $(\mathbf{b}, \mathbf{w})$ satisfying the stated conditions. Since replacing $\mathbf{w}$ by $-\mathbf{w}$ only swaps the two refitted coefficients, we may orient $\mathbf{q} := \mathbf{A}\mathbf{w}/\|\mathbf{A}\mathbf{w}\|$ so that $\mathbf{q}^\top \mathbf{v} \ge 0$. Write

$$\mathbf{q} = c_q \mathbf{v} + \mathbf{q}_\perp, \qquad \mathbf{e} := \mathbf{A}(\mathbf{b} - \bar{\boldsymbol{\beta}}) = e_\parallel \mathbf{v} + \mathbf{e}_\perp,$$

with $\mathbf{q}_\perp, \mathbf{e}_\perp \perp \mathbf{v}$. Then $\|\mathbf{q}_\perp\| \le \delta_{\mathrm{ref}}$ and $\|\mathbf{e}\| \le \delta_{\mathrm{ref}}\Lambda$. In particular, for sufficiently small δ_{ref}, c_q, $1 - \omega - e_\parallel/\Lambda$, and $\omega + e_\parallel/\Lambda$ are bounded away from zero.

By the wedge representation in Fact D.3, the preliminary misassignment events have normalized perturbations

$$\mathbf{h}_1 = \frac{(\mathbf{q}_\perp, 0)}{c_q}, \qquad \mathbf{h}_{2,+} = \frac{(-\mathbf{e}_\perp/\Lambda,\, r)}{1 - \omega - e_\parallel/\Lambda}, \qquad \mathbf{h}_{2,-} = \frac{(\mathbf{e}_\perp/\Lambda,\, -r)}{\omega + e_\parallel/\Lambda}.$$

Hence $\|\mathbf{h}_1\| + \|\mathbf{h}_{2,\pm}\| \le C(\delta_{\mathrm{ref}} + r)$. Thus both misassignment events belong to $\mathcal{W}(s_{\mathrm{pre}})$, where $s_{\mathrm{pre}} := \delta_{\mathrm{ref}} + r$.

The uniform assigned-Gram and refit bounds used in Lemmas D.6 and D.7 therefore apply with s_γ replaced by s_{pre}. On an event of probability at least $1 - Ce^{-t}$,

$$a(\Gamma(\mathbf{b}, \mathbf{w})) \le C\left[s_{\mathrm{pre}}\sqrt{s_{\mathrm{pre}} + \varepsilon_n(t)} + r\sqrt{\frac{p+t}{n}}\right].$$

Choosing δ_{ref}, then r_0, and finally κ sufficiently small makes the right-hand side at most $\rho/2$. □

Proof of Theorem 5.3. Let δ_{ref} be as in Lemma E.10 and set $\delta_0 := \delta_{\mathrm{ref}}$. Choose fixed $0 < A_- < A_+$ such that (45) implies $A_-\Lambda \le \widehat{T} \le A_+\Lambda$, and choose $\delta_1 \in (0, \delta_{\mathrm{ref}})$ sufficiently small. For sufficiently small κ and r_0, Lemma E.2 gives

$$\frac{\|\widehat{\mathbf{b}} - \bar{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}}}{\Lambda} \le \delta_1, \qquad A_-\Lambda \le \widehat{T} \le A_+\Lambda.$$

Apply Lemma E.9 on this neighborhood with tolerance $\eta > 0$. Together with (51),

$$\big\|\widehat{\mathbf{S}}^{-1/2}\widetilde{\mathbf{M}}_U(\widehat{\boldsymbol{\alpha}}, \widehat{T})\widehat{\mathbf{S}}^{-1/2} - \lambda(\widehat{T})\mathbf{v}\mathbf{v}^\top\big\|_{\mathrm{op}} \le \eta + C\delta_1.$$

By (48), $\lambda(\widehat{T}) \ge c$. Set $\widehat{\mathbf{q}} := \widehat{\mathbf{S}}^{1/2}\mathbf{A}\widehat{\mathbf{w}}$. By (52) and $\widehat{\boldsymbol{\Sigma}} = \mathbf{A}\widehat{\mathbf{S}}\mathbf{A}$, $\widehat{\mathbf{q}}$ is a unit leading eigenvector of $\widehat{\mathbf{S}}^{-1/2}\widetilde{\mathbf{M}}_U(\widehat{\boldsymbol{\alpha}}, \widehat{T})\widehat{\mathbf{S}}^{-1/2}$. Hence, for sufficiently small δ_1 and η, the Davis–Kahan theorem Yu et al. (2015) gives

$$\sin\angle(\widehat{\mathbf{q}}, \mathbf{v}) \le \delta_{\mathrm{ref}}/2.$$

On $\mathcal{E}_{\mathrm{G}}(t)$, $\|\widehat{\mathbf{S}}^{-1/2} - \mathbf{I}_p\|_{\mathrm{op}} \le C\sqrt{(p+t)/n}$. Using the representation of the principal angle,

$$\begin{aligned}\sin\angle(\widehat{\mathbf{S}}^{-1/2}\widehat{\mathbf{q}}, \mathbf{v}) &= \frac{\|(\mathbf{I}_p - \mathbf{v}\mathbf{v}^\top)\widehat{\mathbf{S}}^{-1/2}\widehat{\mathbf{q}}\|}{\|\widehat{\mathbf{S}}^{-1/2}\widehat{\mathbf{q}}\|} \\ &\le \frac{\sin\angle(\widehat{\mathbf{q}}, \mathbf{v}) + \|\widehat{\mathbf{S}}^{-1/2} - \mathbf{I}_p\|_{\mathrm{op}}}{1 - \|\widehat{\mathbf{S}}^{-1/2} - \mathbf{I}_p\|_{\mathrm{op}}},\end{aligned}$$

where we used $\widehat{\mathbf{S}}^{-1/2}\widehat{\mathbf{q}} = \widehat{\mathbf{q}} + (\widehat{\mathbf{S}}^{-1/2} - \mathbf{I}_p)\widehat{\mathbf{q}}$ and $\|\widehat{\mathbf{q}}\| = 1$. Since $\mathbf{A}\widehat{\mathbf{w}} = \widehat{\mathbf{S}}^{-1/2}\widehat{\mathbf{q}}$ and $\mathbf{A}\boldsymbol{\theta} = \Lambda\mathbf{v}$, for sufficiently small κ,

$$\sin\angle(\mathbf{A}\widehat{\mathbf{w}}, \mathbf{A}\boldsymbol{\theta}) \le \delta_{\mathrm{ref}} = \delta_0.$$

Together with the pooled-coefficient bound, this shows that $(\widehat{\mathbf{b}}, \widehat{\mathbf{w}})$ satisfies the conditions of Lemma E.10. Since $\gamma^{(0)} = \Gamma(\widehat{\mathbf{b}}, \widehat{\mathbf{w}})$,

$$a(\gamma^{(0)}) \le \rho/2.$$

The stated probability follows by intersecting the preceding events. □

E.5 Completion of efficient recovery

Proof of Theorem 1.2. Take $r_0 \le 1$ sufficiently small so that Theorem 5.3 and Corollary 4.3 apply for $r \le r_0$. By Theorem 5.3,

$$a(\gamma^{(0)}) \le \rho/2$$

with the required probability. Since the bounds in Corollary 4.3 hold uniformly over all starting pairs satisfying this condition, they apply to the random pair $\gamma^{(0)}$ on the initialization event. Accounting for the failure probability of this event yields the expected mis-clustering, exact-recovery, and high-probability bounds in Theorem 1.2, for sufficiently small κ_2 and suitably adjusted constants.

For $r > r_0$, the three bounds follow from $\bar{h} \le 1$ and probabilities being at most one after enlarging the constants. □

Proof of Corollary 1.3. For sufficiently small κ, Theorems 5.3 and 4.1 may both be applied with $r = 0$ and $t = cn$ for a sufficiently small constant $c > 0$. Intersecting their events gives an event $\mathcal{G}$ with $\mathbb{P}(\mathcal{G}^c) \le Ce^{-cn}$ on which $a(\gamma^{(0)}) \le \rho/2$. Since $Q^\circ = 0$ when $r = 0$, the recursion in the proof of Theorem 4.1 and $\mathbb{E}H^{\mathrm{w}}(\gamma^{(0)}) \le 1$ give

$$\mathbb{E}\big[H^{\mathrm{w}}(\gamma^{(K)})\mathbf{I}(\mathcal{G})\big] \le 2^{-K}.$$

Applying the same margin-decomposition argument as in the proof of Corollary 4.3, and using $\widehat{\mathbf{z}} = \mathbf{z}^{(K+1)}$, for every $h > 0$,

$$\mathbb{P}(\widehat{\mathbf{z}} \notin \{\mathbf{z}^*, -\mathbf{z}^*\}) \le Ce^{-cn} + nh + \frac{n2^{-K}}{h^2}.$$

Taking $h = 2^{-K/3}$ yields

$$\mathbb{P}(\widehat{\mathbf{z}} \notin \{\mathbf{z}^*, -\mathbf{z}^*\}) \le Ce^{-cn} + 2n2^{-K/3}.$$

The two conclusions follow for $K \ge C \log n$ and $K \ge Cn$, respectively, after adjusting the constants. □

E.6 Proof of Lemma E.8

We first establish concentration on a finite product net, then control empirical and population interpolation between neighboring parameter values, and finally combine these bounds.

Step 1. Pointwise concentration and product net.

Define

$$\widehat{\mathbf{M}}_U(\boldsymbol{\alpha}, T) := \frac{1}{n}\sum_{i=1}^{n} h_T(y_i - \mathbf{u}_i^\top \boldsymbol{\alpha})(\mathbf{u}_i\mathbf{u}_i^\top - \mathbf{I}_p).$$

Fix $(\boldsymbol{\alpha}, T)$ and write $h_i := h_T(y_i - \mathbf{u}_i^\top \boldsymbol{\alpha})$. For every unit vector $\mathbf{q}$,

$$\mathbf{q}^\top\big(\widehat{\mathbf{M}}_U(\boldsymbol{\alpha}, T) - \mathbf{M}_U(\boldsymbol{\alpha}, T)\big)\mathbf{q} = \frac{1}{n}\sum_{i=1}^{n}\Big[h_i\big((\mathbf{q}^\top\mathbf{u}_i)^2 - 1\big) - \mathbb{E}\big[h_i\big((\mathbf{q}^\top\mathbf{u}_i)^2 - 1\big)\big]\Big].$$

Since $0 \le h_i \le 1$, the summands are independent, centered, and sub-exponential with uniformly bounded sub-exponential norms. Bernstein's inequality and a standard sphere-covering argument (see, e.g., Vershynin, 2018, Corollary 4.2.13 and Exercise 4.4.2) therefore give, for every $s \ge 1$, with probability at least $1 - Ce^{-s}$,

$$\|\widehat{\mathbf{M}}_U(\boldsymbol{\alpha}, T) - \mathbf{M}_U(\boldsymbol{\alpha}, T)\|_{\mathrm{op}} \le C\left(\sqrt{\frac{p+s}{n}} + \frac{p+s}{n}\right).$$

Fix $\zeta \in (0,1)$. Let $\mathcal{N}_\alpha$ be a $\zeta\Lambda$-net of $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\| \le \delta\Lambda\}$, and let $\mathcal{N}_T$ be a ζ-net of $[\log A_-, \log A_+]$ for $\log(T/\Lambda)$. Then

$$|\mathcal{N}_\alpha|\,|\mathcal{N}_T| \le \exp(C_\zeta p).$$

Applying the pointwise bound on the product net with $s = t + C_\zeta p$ and taking a union bound gives, with probability at least $1 - Ce^{-t}$,

$$\max_{(\boldsymbol{\alpha}_0, T_0) \in \mathcal{N}_\alpha \times \mathcal{N}_T} \|\widehat{\mathbf{M}}_U(\boldsymbol{\alpha}_0, T_0) - \mathbf{M}_U(\boldsymbol{\alpha}_0, T_0)\|_{\mathrm{op}} \le C_\zeta\left(\sqrt{\frac{p+t}{n}} + \frac{p+t}{n}\right). \tag{53}$$

Step 2. Empirical interpolation.

Suppose $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| \le \zeta\Lambda$ and, when the two coefficients differ, set $\mathbf{e} := (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)/\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|$. Since $T \ge A_-\Lambda$, Fact E.4 gives

$$|h_T(y_i - \mathbf{u}_i^\top \boldsymbol{\alpha}) - h_T(y_i - \mathbf{u}_i^\top \boldsymbol{\alpha}_0)| \le C\min(1, \zeta|\mathbf{u}_i^\top \mathbf{e}|).$$

Hence, for every unit vector $\mathbf{q}$,

$$\Big|\mathbf{q}^\top\big(\widehat{\mathbf{M}}_U(\boldsymbol{\alpha}, T) - \widehat{\mathbf{M}}_U(\boldsymbol{\alpha}_0, T)\big)\mathbf{q}\Big| \le \frac{C}{n}\sum_{i=1}^{n}\min(1, \zeta|\mathbf{u}_i^\top \mathbf{e}|)\big((\mathbf{q}^\top\mathbf{u}_i)^2 + 1\big).$$

Fix $L \ge 2$ and define $I_{\mathbf{a},L} := \{i : |\mathbf{u}_i^\top \mathbf{a}| > L\}$ and $\tau_L := \min\big(1, e^{-L^2/2} + \sqrt{\psi_n(p,t)}\big)$. The class of complements of centered slabs has VC dimension $O(p)$, so Lemma A.1(i) gives $\sup_{\|\mathbf{a}\|=1} |I_{\mathbf{a},L}|/n \le$

$C\tau_L$, with probability at least $1 - e^{-t}$. On the event in Lemma A.2, parts (i)–(ii) then give, uniformly over unit $\mathbf{a}, \mathbf{q}$,

$$\frac{1}{n}\sum_{i=1}^{n}\big((\mathbf{q}^\top \mathbf{u}_i)^2 + 1\big) \le C, \qquad \frac{1}{n}\sum_{i \in I_{\mathbf{a},L}}\big((\mathbf{q}^\top \mathbf{u}_i)^2 + 1\big) \le C\bigg(\frac{p+t}{n} + \tau_L \log\frac{e}{\tau_L}\bigg).$$

Using $\min(1, \zeta|\mathbf{u}_i^\top \mathbf{e}|) \le \zeta L$ on $I^c_{\mathbf{e},L}$, and $\min(1, \zeta|\mathbf{u}_i^\top \mathbf{e}|) \le 1$ on $I_{\mathbf{e},L}$, gives

$$\|\widehat{\mathbf{M}}_U(\boldsymbol{\alpha}, T) - \widehat{\mathbf{M}}_U(\boldsymbol{\alpha}_0, T)\|_{\mathrm{op}} \le C\bigg(\zeta L + \frac{p+t}{n} + \tau_L \log\frac{e}{\tau_L}\bigg).$$

For the scale parameter, Fact E.4 gives $|h_T(x) - h_{T_0}(x)| \le C\zeta$ whenever $|\log(T/T_0)| \le \zeta$. Therefore,

$$\|\widehat{\mathbf{M}}_U(\boldsymbol{\alpha}, T) - \widehat{\mathbf{M}}_U(\boldsymbol{\alpha}, T_0)\|_{\mathrm{op}} \le C\zeta.$$

Step 3. Population interpolation.

By (50) in Fact E.7, if $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| \le \zeta\Lambda$ and $|\log(T/T_0)| \le \zeta$, then

$$\|\mathbf{M}_U(\boldsymbol{\alpha}, T) - \mathbf{M}_U(\boldsymbol{\alpha}_0, T_0)\|_{\mathrm{op}} \le C\zeta.$$

Step 4. Completion.

For each $(\boldsymbol{\alpha}, T)$ in the parameter neighborhood, choose $(\boldsymbol{\alpha}_0, T_0) \in \mathcal{N}_\alpha \times \mathcal{N}_T$ satisfying the preceding two bounds. Combining (53) with the empirical and population interpolation bounds gives

$$\|\widehat{\mathbf{M}}_U(\boldsymbol{\alpha}, T) - \mathbf{M}_U(\boldsymbol{\alpha}, T)\|_{\mathrm{op}} \le C_\zeta\bigg(\sqrt{\frac{p+t}{n}} + \frac{p+t}{n}\bigg) + C\bigg(\zeta L + \frac{p+t}{n} + \tau_L \log\frac{e}{\tau_L} + \zeta\bigg).$$

Choose L so that $Ce^{-L^2/2}(1 + L^2) \le \eta/4$, and then take $\zeta > 0$ with $C\zeta(L+1) \le \eta/4$. Under $p + t + \log(en) \le \kappa n$, both $(p+t)/n$ and $\psi_n(p,t)$ tend uniformly to zero as $\kappa \to 0$. Hence, for sufficiently small κ, the remaining terms are at most $\eta/2$. These bounds hold simultaneously with probability at least $1 - Ce^{-t}$, which proves the claim.

F Implementation details and additional experiments

This appendix provides implementation details and additional simulations for the experiments in Section 6. Unless stated otherwise, we use the same simulation setting.

Reference design. To generate a covariance matrix with condition number $\kappa_{\boldsymbol{\Sigma}}$, let $\mathbf{O}$ be the orthogonal factor in the QR decomposition of a $p \times p$ matrix with independent $N(0,1)$ entries and set

$$\boldsymbol{\Sigma} = \mathbf{O}\,\mathrm{diag}\left(\kappa_{\boldsymbol{\Sigma}}^{-1/2+(j-1)/(p-1)}\right)_{j=1}^{p}\mathbf{O}^\top.$$

We choose

$$\mathbf{v} = \frac{(p, p-1, \ldots, 1)^\top}{\|(p, p-1, \ldots, 1)\|_2}, \qquad \boldsymbol{\theta} = \Lambda\boldsymbol{\Sigma}^{-1/2}\mathbf{v}.$$

To ensure the orthogonality condition $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}\boldsymbol{\theta} = 0$, we set

$$\mathbf{v}_\mu = \frac{(\mathbf{I}_p - \mathbf{v}\mathbf{v}^\top)\mathbf{e}_p}{\|(\mathbf{I}_p - \mathbf{v}\mathbf{v}^\top)\mathbf{e}_p\|_2}, \qquad \boldsymbol{\mu} = 2\Lambda\boldsymbol{\Sigma}^{-1/2}\mathbf{v}_\mu.$$

When p or $\kappa_{\boldsymbol{\Sigma}}$ changes, we regenerate $\boldsymbol{\Sigma}$, $\boldsymbol{\theta}$, and $\boldsymbol{\mu}$ according to this construction.

We next describe the implementation details for the comparison methods.

Initializers followed by hard least squares.

- **Proposed.** We use the initialization in Algorithm 3.1.

- **SVD-grid.** We adapt the initializer of Yi et al. (2014) to general $\mathbf{\Sigma}$ using empirical whitening. In the whitened coordinates, we search the leading two-dimensional eigenspace on a grid with angular spacing 0.3 and select the coefficient pair with the smallest RSS. Because the component coefficients are not normalized in our setting, we set the search radius to the empirical response scale $(n^{-1}\sum_i y_i^2)^{1/2}$.

- **Random-start initialization.** We generate 100 independent Bernoulli partitions, discard those with fewer than p observations in either group, fit separate regressions within the two groups for each remaining partition, and retain the partition with the smallest RSS.

- **Spectral Experts.** We implement the moment method of Chaganty & Liang (2013), estimating the first three coefficient moments and recovering the two regression components by spectral and tensor decomposition. To avoid explicitly forming the tensor-product design matrices associated with $\mathbf{x}_i \otimes \mathbf{x}_i$ and $\mathbf{x}_i^{\otimes 3}$, we implement the second and third-order moment regressions through rank-two factorizations optimized by L-BFGS-B. This factorized implementation is nonconvex; as a check, we also run the original convex moment regressions in lower-dimensional settings, where the two implementations give similar recovery results. We allow at most 500 L-BFGS-B iterations with tolerance 10^{-7}.

  For comparability with the other methods, we use the full sample for all three moment regressions, rather than the separate samples used in the theory. The noise variance is estimated as the nonnegative intercept in the second-moment regression, except in the group-specific intercept experiment, where it is not separately identifiable and the true simulation variance is used. The recovered regression coefficients are then used to initialize hard least squares.

- **Convex formulation.** We adapt the estimator of Chen et al. (2014) using empirical whitening and solve the resulting convex program by FISTA. Since the method requires the noise variance, we use the estimate from the Spectral Experts second-moment regression. We set

$$\widehat{\lambda}_{\mathrm{CY}} = \widehat{\sigma}_{\mathrm{SE}} \left[2\left(\frac{1}{n}\sum_i y_i^2 - \widehat{\sigma}_{\mathrm{SE}}^2 \right)_+^{1/2} + \widehat{\sigma}_{\mathrm{SE}} \right] \sqrt{\frac{p}{n}},$$

  where $\widehat{\sigma}_{\mathrm{SE}}^2$ is the noise-variance estimate given by the nonnegative intercept in the Spectral Experts second-moment regression. We take the unspecified multiplicative constant in the theoretical regularization level to be one. FISTA is run for at most 4000 iterations with tolerance 10^{-5}.

Clusterwise optimization. For multi-start Späth (Späth, 1979), we run the one-observation exchange algorithm from 10 feasible Bernoulli initial partitions and retain the solution with the smallest RSS. For GA–Lloyd (Park et al., 2017), each generation maintains 10 candidate coefficient pairs; we use mutation probability 0.01 and stop after 50 generations without improvement. Each Späth exchange search terminates when no admissible single-observation move decreases the RSS.

Concave fusion. We implement the estimator of Ma et al. (2020) with either its original ridge-fusion initialization or the proposed initialization, using the same MCP tuning grid for both. We fix the target number of groups at two. Among solutions with two fitted groups, we select the one with the smallest RSS; if none is available, we take the nearest over-segmented solution and merge the closest coefficient centers to obtain a two-group partition. We set the MCP concavity parameter to 3 and the augmented-Lagrangian penalty parameter in ADMM to 1. ADMM is run for at most 100 iterations, with absolute and relative tolerances 10^{-5} and 10^{-4}, respectively. Since the original ridge-fusion initialization performs worse than the proposed initialization, we omit it from the appendix figures.

F.1 Additional comparisons

At fixed $r = 0.1$, Figure 6 varies p, n, $\kappa_{\mathbf{\Sigma}}$, and $\|\boldsymbol{\mu}\|_{\mathbf{\Sigma}}/\Lambda$ one at a time, with the remaining parameters fixed at their reference values.

Figure 6 shows similar mis-clustering error for the proposed method across the covariance condition numbers and common-center magnitudes considered, whereas Spectral Experts and the convex initializer become less accurate as the common center grows.

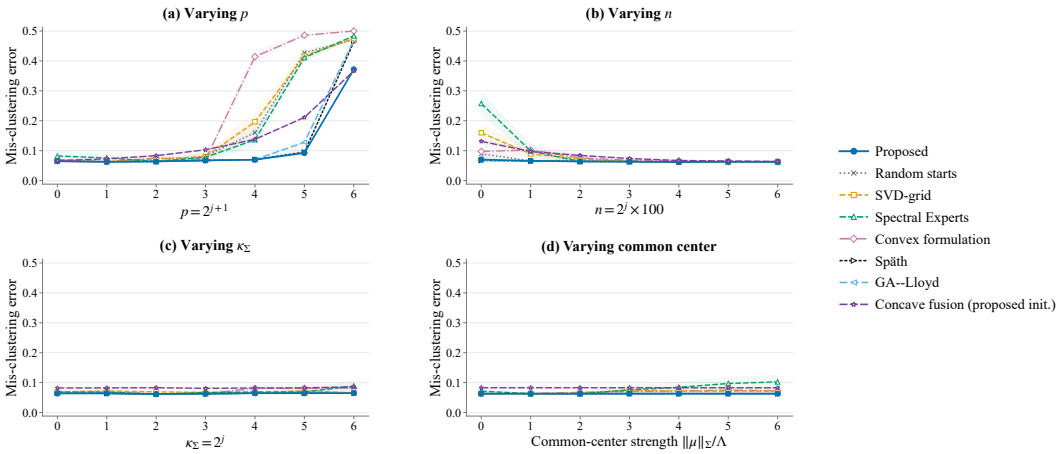


Figure 6: Additional comparisons at fixed $r = 0.1$.

F.2 Running time

Running times were measured on an AMD Ryzen 7 5800H processor, with each method run in a single process and the numerical-library thread count fixed at one.

Table 2: Mean running time in seconds for the noiseless experiment.

(a) Varying n

Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$	$n = 3200$	$n = 6400$
Proposed	**0.0007**	**0.0007**	**0.0008**	**0.0010**	**0.0013**	**0.0019**	**0.0034**
Random starts	0.0075	0.0077	0.0084	0.0099	0.0125	0.0174	0.0270
SVD-grid	0.0014	0.0015	0.0018	0.0023	0.0032	0.0052	0.0101
Spectral Experts	0.2578	0.2141	0.1927	0.2007	0.2384	0.3173	0.5957
Convex formulation	0.9222	0.5661	0.4699	0.4625	0.4573	0.5410	0.6703
Späth	0.2042	0.3657	0.7265	1.4024	2.7644	5.3885	10.7358
GA–Lloyd	0.0820	0.0867	0.1034	0.1322	0.1928	0.3057	0.5470
Concave fusion	0.0065	0.0390	0.1215	0.6853	2.8099	8.0854	41.7158
Concave fusion (proposed init.)	0.0018	0.0029	0.0044	0.0157	0.0615	0.2551	1.0757

(b) Varying p

Method	$p = 2$	$p = 4$	$p = 8$	$p = 16$	$p = 32$	$p = 64$	$p = 128$
Proposed	**0.0005**	**0.0006**	**0.0008**	**0.0011**	**0.0026**	**0.0097**	**0.0442**
Random starts	0.0075	0.0078	0.0084	0.0097	0.0152	0.0335	0.0915
SVD-grid	0.0011	0.0014	0.0018	0.0028	0.0061	0.0186	0.0485
Spectral Experts	0.1729	0.1706	0.1927	0.3067	0.5261	1.0484	2.8505
Convex formulation	0.1324	0.2848	0.4699	1.0576	1.1488	0.2511	0.2704
Späth	0.5648	0.6736	0.7265	0.8802	1.3192	2.9462	5.9383
GA–Lloyd	0.0839	0.0911	0.1034	0.1462	0.3055	0.9453	6.2251
Concave fusion	0.0070	0.0391	0.1215	0.3329	0.5245	0.1663	0.1930
Concave fusion (proposed init.)	0.0013	0.0025	0.0044	0.0238	0.1319	0.2630	0.2878

Tables 2 and 3 report all methods used in the noiseless and noisy experiments, respectively. For concave fusion, we report the ADMM time at the selected tuning parameter, excluding the cost of the remaining regularization path.

The proposed method and SVD-grid have comparable running times and are substantially faster than the other methods. Spectral Experts and the convex formulation incur additional iterative estimation or optimization, while Späth, GA–Lloyd, and concave fusion become more costly as the sample size increases. The proposed initialization also reduces the ADMM time of concave fusion at the selected tuning parameter in most settings.

Table 3: Mean running time in seconds for the Gaussian-noise experiment.

(a) Varying n

Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$	$n = 3200$	$n = 6400$
Proposed	**0.0012**	**0.0009**	**0.0010**	**0.0015**	**0.0018**	**0.0028**	**0.0044**
Random starts	0.0087	0.0156	0.0094	0.0111	0.0152	0.0187	0.0304
SVD-grid	0.0016	0.0019	0.0021	0.0031	0.0047	0.0063	0.0111
Spectral Experts	0.2797	0.1963	0.1935	0.1805	0.2153	0.2949	0.4860
Convex formulation	0.8060	0.5350	0.4258	0.4107	0.4339	0.4868	0.6253
Späth	0.2201	0.3918	0.7780	1.5848	3.3189	6.9470	13.6481
GA–Lloyd	0.0866	0.0911	0.1127	0.1451	0.2283	0.4061	0.7158
Concave fusion	0.0068	0.0505	0.1760	0.7977	3.0580	7.9527	41.1869
Concave fusion (proposed init.)	0.0027	0.0075	0.0211	0.0173	0.1701	0.2738	1.0629

(b) Varying p

Method	$p = 2$	$p = 4$	$p = 8$	$p = 16$	$p = 32$	$p = 64$	$p = 128$
Proposed	**0.0007**	**0.0008**	**0.0010**	**0.0014**	**0.0049**	**0.0143**	**0.0488**
Random starts	0.0099	0.0087	0.0094	0.0109	0.0279	0.0351	0.1006
SVD-grid	0.0013	0.0017	0.0021	0.0031	0.0112	0.0200	0.0535
Spectral Experts	0.1384	0.1848	0.1935	0.2937	0.4674	0.7714	2.0569
Convex formulation	0.1285	0.2448	0.4258	0.9850	1.0929	0.2325	0.2382
Späth	0.5800	0.6998	0.7780	1.0297	1.5771	3.1205	5.9347
GA–Lloyd	0.0919	0.0992	0.1127	0.1690	0.3544	1.3576	5.8839
Concave fusion	0.0217	0.0450	0.1760	0.3863	0.4343	0.1523	0.2267
Concave fusion (proposed init.)	0.0021	0.0030	0.0211	0.0819	0.2386	0.2430	0.2937

F.3 Bounded residual weighting

To isolate the effect of bounded residual weighting, we compare the spectral initializer in Algorithm 3.1 with the same construction using $\widehat{e}_i^2$ in place of $h_{\widehat{T}}(\widehat{e}_i)$, where $\widehat{e}_i = y_i - \mathbf{x}_i^\top \widehat{\mathbf{b}}$. Panel (a) of Figure 7 reports $\sin\angle(\boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{w}}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta})$ before label assignment and hard least-squares refinement, varying p/n at the reference setting. Panels (b)–(e) use the four settings in Figure 4.

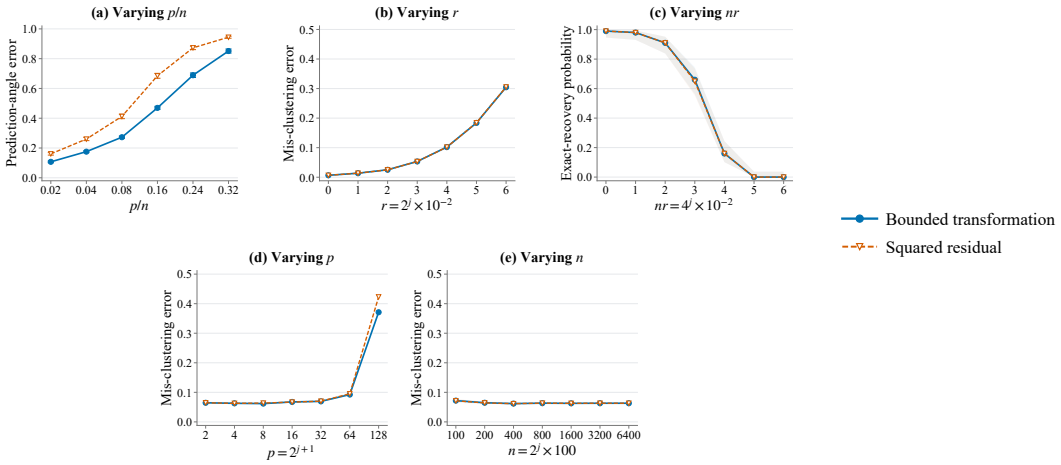


Figure 7: Bounded versus squared-residual weighting in the spectral initializer.

Bounded weighting yields more accurate spectral directions over the dimensions considered. After hard least-squares refinement, the two initializers have similar recovery performance except at the largest dimensions.

F.4 Unequal group sizes

Figure 8 repeats the four comparisons in Figure 4 under a 1:3 group-size allocation. The sample-size panel preserves this proportion.

Recovery deteriorates with increasing dimension earlier than in the balanced setting because the smaller group contains fewer observations for groupwise least-squares fitting.

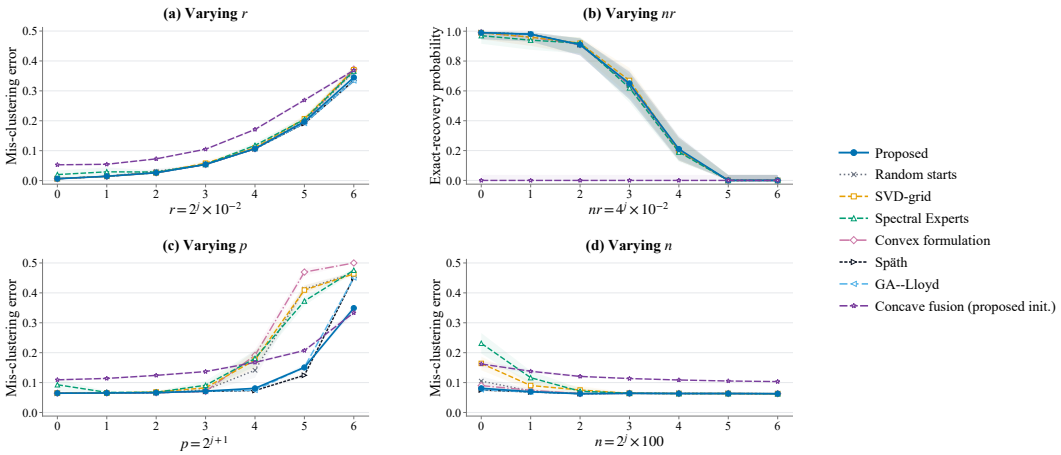


Figure 8: Recovery with a 1:3 group-size ratio.

F.5 Group-specific intercepts

We add group-specific intercepts and apply all methods to the augmented regression design

$$y_i = a^*_{z_i^*} + \mathbf{x}_i^\top \boldsymbol{\beta}^*_{z_i^*} + \epsilon_i = \widetilde{\mathbf{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*_{z_i^*} + \epsilon_i, \qquad \widetilde{\mathbf{x}}_i = (1, \mathbf{x}_i^\top)^\top.$$

Let $\Delta a = a^*_+ - a^*_-$ and define the total separation by

$$\Lambda^2_{\text{aff}} = (\Delta a)^2 + \Lambda^2.$$

We set $\Lambda_{\text{aff}} = 1$, with equal contributions from the intercept and slope differences,

$$a^*_+ = -a^*_- = \frac{1}{2\sqrt{2}}, \qquad \Lambda = \frac{1}{\sqrt{2}},$$

and take $\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} = 2\Lambda_{\text{aff}}$. With the remaining parameters fixed at their reference values, Figure 9 varies the noise level on the $\sigma/\Lambda_{\text{aff}}$ and $n\sigma/\Lambda_{\text{aff}}$ scales.

The SVD-grid, Spectral Experts, and convex initializers use empirical whitening of the augmented design. With the constant covariate, the intercept in the Spectral Experts second-moment regression is not separately identifiable from the noise variance, so we use the true noise variance for Spectral Experts. For the convex initializer, we retain the nonnegative second-moment intercept as the noise-variance input.

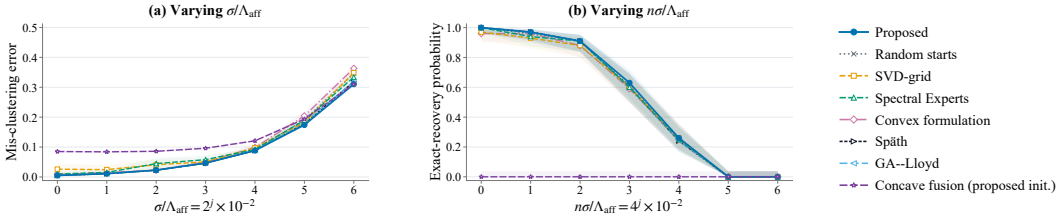


Figure 9: Recovery with group-specific intercepts.

The qualitative pattern is similar to that without group-specific intercepts. For comparison with the no-intercept model, when the regression components are known, the difference between the two possible response means satisfies

$$\Delta a + \mathbf{x}^\top \boldsymbol{\theta} \sim N(\Delta a, \Lambda^2).$$

The corresponding single-membership Bayes error is

$$q_{\text{aff}} = \mathbb{E}[\Phi(-|\Delta a + \mathbf{x}^\top \boldsymbol{\theta}|/(2\sigma))] \sim \frac{2}{\pi}\frac{\sigma}{\Lambda} \exp(-(\Delta a)^2/(2\Lambda^2)),$$

as $\sigma/\Lambda \to 0$. Thus, for fixed $\Delta a/\Lambda$, the oracle mis-clustering scale is σ/Λ, while the exact-recovery threshold is $n\sigma/\Lambda \to 0$.

F.6 Error distributions and dependence

We consider independent Laplace and Student-t_5 errors, each scaled to variance σ^2, as well as Gaussian AR(1) errors with

$$\mathrm{Cov}(\epsilon_i, \epsilon_j) = \sigma^2 (0.5)^{|i-j|}.$$

The covariates are Gaussian with the reference covariance $\mathbf{\Sigma}$ and are independent of the errors.

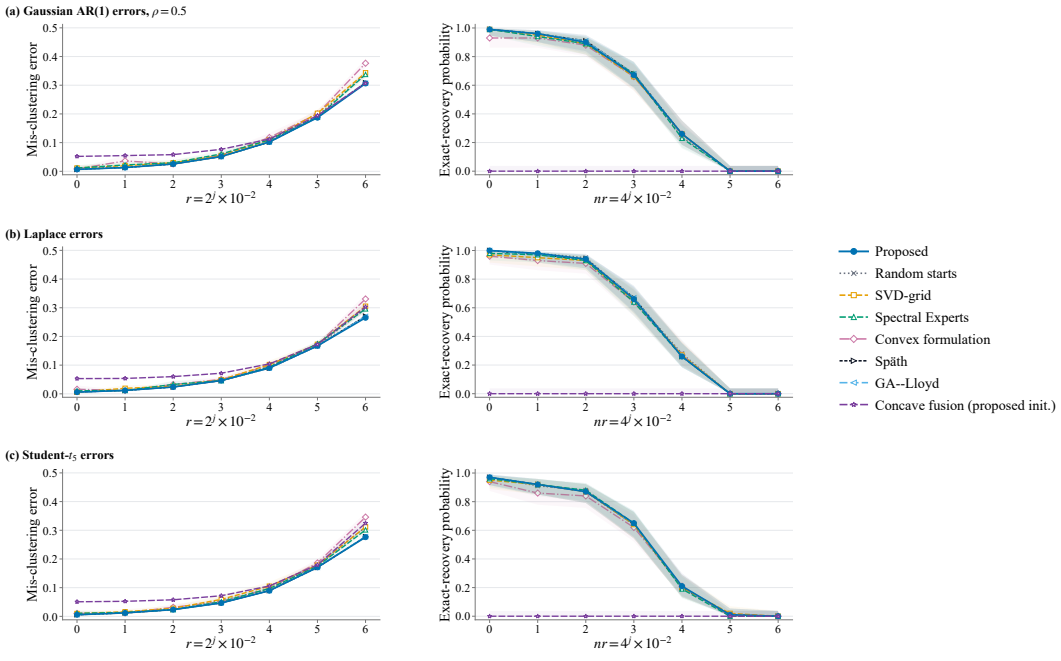


Figure 10: Recovery under Gaussian AR(1), Laplace, and Student-t_5 errors, each with marginal variance σ^2. Columns report mis-clustering error and exact recovery.

Figure 10 shows qualitative recovery transitions similar to those under independent Gaussian errors.

F.7 Alternative covariate distributions

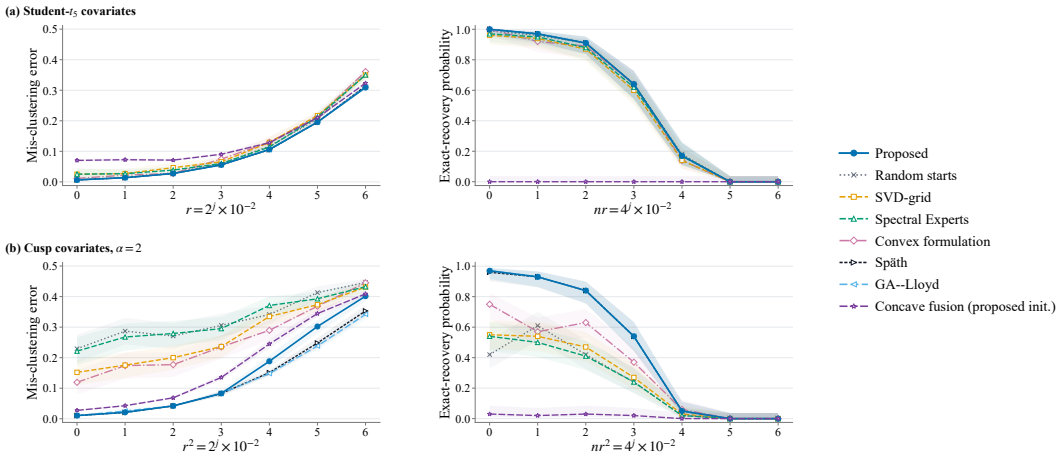


Figure 11: Recovery under Student-t_5 and $\alpha = 2$ covariate designs.

We generate $\mathbf{x}_i = \mathbf{\Sigma}^{1/2}\mathbf{u}_i$, where the coordinates of $\mathbf{u}_i$ are independent, mean zero, and variance one. We consider standardized Student-t_5 coordinates and coordinates with density $f(u) = |u|/2$

on $[-\sqrt{2}, \sqrt{2}]$, while retaining Gaussian errors. For the latter design, we set $\mathbf{v} = \mathbf{e}_1$, so that $\mathbf{x}_i^\top \boldsymbol{\theta}/\Lambda = u_{i1}$ and

$$\mathbb{P}\big(|\mathbf{x}_i^\top \boldsymbol{\theta}|/\Lambda \le t\big) = \mathbb{P}(|u_{i1}| \le t) = t^2/2, \qquad 0 \le t \le \sqrt{2}.$$

This gives quadratic small-ball behavior near zero. We refer to this as the $\alpha = 2$ design. The Student-t_5 panels use the r and nr scales, whereas the $\alpha = 2$ panels use r^2 and nr^2. Figure 11 reports the corresponding recovery results.

The $\alpha = 2$ experiment exhibits recovery transitions on the squared scales suggested by its quadratic small-ball probability, illustrating how the mass of the random margin near zero determines the recovery scale.